



APLICACIONES DE MARCADORES GENÉTICOS EN LA INFERENCIA DE
CARACTERÍSTICAS VISIBLES EXTERNAS Y ORIGEN ANCESTRAL CON
FINES FORENSES

FDP: "FORENSIC DNA PHENOTYPING"

YARIMAR RUIZ OROZCO

FDP: "FORENSIC DNA PHENOTYPING"

APLICACIONES DE MARCADORES GENÉTICOS
EN LA INFERENCIA DE CARACTERÍSTICAS
VISIBLES EXTERNAS Y ORIGEN ANCESTRAL
CON FINES FORENSES



FACULTAD DE MEDICINA
DEPARTAMENTO DE
ANATOMÍA PATOLÓGICA
Y CIENCIAS FORENSES

YARIMAR RUIZ OROZCO
2012

Universidad Santiago de Compostela

Facultad de Medicina

Departamento de Anatomía Patológica y Ciencias Forenses



FDP: “Forensic DNA Phenotyping”

**Aplicaciones de marcadores genéticos en la inferencia
de características visibles externas y origen ancestral
con fines forenses**

Memoria que presenta para optar al grado de doctor,

Yarimar Ruiz Orozco

En Santiago de Compostela, Junio 2012



La Doctora María Victoria Lareu Huidobro, y el Doctor Ángel Carracedo Álvarez, Catedráticos del departamento de Anatomía Patológica y Ciencias Forenses de la Facultad de Medicina de la Universidad de Santiago de Compostela,

CERTIFICAN:

Que la presente memoria que lleva por título **FDP: “Forensic DNA Phenotyping”: Aplicaciones de marcadores genéticos en la inferencia de características visibles externas y origen ancestral con fines forenses**, de la licenciada en biología por la Universidad de Central de Venezuela, *Yarimar Ruiz Orozco*, ha sido realizada bajo nuestra dirección, considerándola en condiciones para optar al grado de doctor y autorizándola para su presentación ante el tribunal correspondiente.

Y para que así conste, firmamos la presente en Santiago de Compostela, a 8 de junio de 2012.

Prof. Dr. Ángel Carracedo Álvarez

Prof. Dra. María Victoria Lareu Huidobro

Yarimar Ruiz Orozco

Este trabajo ha sido financiado en parte con la beca otorgada por la Fundación Gran Mariscal de Ayacucho (FUNDAYACUCHO) perteneciente al convenio “Robinson y Freire: Hacia la educación popular” E-228-585-2001, así como por los proyectos otorgados por la Xunta de Galicia: INCITE 09 209163PPR y PGIDTI0T6P-XIB228195PR

Indice

<i>Dedicatoria</i>	11
<i>Agradecimientos</i>	13
<i>Abreviaturas</i>	15
Introducción	18
I. Evolución de la genética forense y sus herramientas	19
I.1. Single Nucleotide Polymorphisms (SNPs)	21
<i>I.1.a. Tecnologías para la detección y genotipado de SNPs</i>	<i>22</i>
<i>I.1.b. Bases de datos de SNPs</i>	<i>29</i>
<i>I.1.c. Aplicaciones forenses de los SNPs</i>	<i>30</i>
II. La predicción de rasgos físicos en un contexto forense	35
II.1. Marcadores informativos de grupos ancestrales (AIMs)	36
<i>II.1.a. Estratificación y mezcla de poblaciones</i>	<i>37</i>
<i>II.1.b. Grupo Ancestral Biogeográfico (BGA)</i>	<i>43</i>
<i>II.1.c. Determinación de Linajes</i>	<i>44</i>
<i>II.1.d. Métodos estadísticos para inferir BGA, estructura y mezcla de poblaciones</i>	<i>48</i>
<i>II.1.e. Análisis de genomas completos en poblaciones humanas</i>	<i>50</i>
II.2. External Visible Characteristics (EVC)	51
<i>II.2.a. Quantitative Trait Locus (QTL)</i>	<i>51</i>
<i>II.2.b. Tecnologías para la identificación de genes asociados a rasgos complejos</i>	<i>54</i>
<i>II.2.c. Algunos EVCs estudiados con aplicación forense</i>	<i>56</i>
<i>II.2.d. La pigmentación humana</i>	<i>59</i>

II.2.d.1. Patrones de distribución de la pigmentación humana a nivel mundial 60

II.2.d.2. Melanogénesis 63

II.2.d.3. Pigmentación del iris 66

II.2.d.4. Investigación genética del color de ojos 69

*II.2.e. Consideraciones éticas y legales sobre la inferencia de EVCs a partir del ADN
78*

II.3. Algunos casos aplicados al estudio de AIMs 80

II.3.a. Operación Minstead 80

II.3.b. Ataque terrorista en Madrid 11-M 81

Justificación y Objetivos84

III. Justificación 85

IV. Objetivos 86

Resultados y Discusión88

V. Preámbulo 89

**V.1. Analysis of the SNPforID 52-plex markers in four Natives
American populations from Venezuela 90**

**V.2. Development of a panel of genome-wide ancestry informative
markers to study admixture throughout the Americas 95**

**V.3. A Melting pot of multicontinental mtDNA lineages in
admixed venezuelans 112**

V.4. PIMA: A population indicative multiplex for the Americas 123

Bloque 1- Discusión 145

V.5. Further development of forensic eye colour predictive tests 149

V.6. A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type 163

V.7. A researcher's guide to STRUCTURE software: applications, parameter settings and supporting software 166

V.8. Assessing the forensic potential of an eye colour predictive test in challenging DNA 194

Bloque 2- Discusión 201

Conclusiones207

Apéndice 213

Bibliografía 233

Dedicatoria

A Cecilia, Joxelyn, Jeiny, Gindy y Mary.

A Ezequiel.

Agradecimientos

Un trabajo feito en equipo:

La culminación de este trabajo no hubiese sido posible sin el aporte de tantas personas que estuvieron presentes durante esta experiencia de cinco años.

Ante todo, agradezco a mi madre y hermanas por el gran apoyo que siempre me han brindado en cada una de mis decisiones, por haberme inculcado un espíritu de lucha y de motivación cada día, a pesar de la distancia. Ezequiel, gracias por acompañarme en este viaje transoceánico lleno de vivencias tan enriquecedoras! A ustedes les dedico este y todos mis logros!

Debo reconocer como extranjera, que hace cinco años desconocía de la existencia de esta ciudad en el mapa!. Al acabar la licenciatura y mientras buscaba el rumbo, tuve la suerte de conversar una mañana con quien fuera padrino de mi promoción y contarle lo que me gustaría hacer en un futuro: genética forense. El, no dudó ni un segundo en recomendarme un grupo de investigación en “Santiago de Compostela”, donde podría tener la oportunidad de formarme junto a grandes profesionales expertos en el área. Así que Dr. Ramirez, muchas gracias por tan acertado consejo!.

*Al llegar a Santiago, me recibió una ciudad acogedora y una universidad con más años de historia que mi propio país! Luego, en el departamento me sorprendió ver a tantas personas trabajando juntas en un espacio algo confinado, entonces, ahí comprendí que el mayor recurso que tiene este grupo en definitiva **es su gente**. Gracias Ángel y Maviky por haberme permitido formar parte de este equipo y por haber confiado en mi. Sé que han venido tiempos difíciles, pero ojalá ambos continúen esta la gran labor de enseñar no sólo dentro de las aulas, sino de seguir dando la oportunidad a más personas de aprender, formarse y trabajar en vuestro equipo.*

A ustedes, “chic@s del lab”, de los que todos los días aprendo algo nuevo, gracias por compartir tantas horas (creo que he pasado más horas con ustedes que con cualquiera durante estos años!), y por tratar de mantener siempre un buen ambiente, algo difícil entre personas distintas pero a la vez, tan auténticas. Esto además incluye al personal de Legal, la Fundación y el CIBERER, quienes siempre han sido colaboradores y de gran ayuda. Extiendo también estos agradecimientos para los que ya no están trabajando en el grupo, a quienes les guardo un especial cariño. Ha sido un placer compartir con ustedes la co-autoría de los trabajos de investigación que fueron realizados...y los que quedan pendientes! Thanks Chris for be so helpful and kind during these years!

Finalmente, quiero agradecer a los “ciudadanos del mundo” que he conocido viviendo aquí en Santiago, quienes han hecho de esta estancia una experiencia aún más amena, llenando de alegría cada momento compartido y a los cuales les deseo mucha suerte con sus planes futuros!

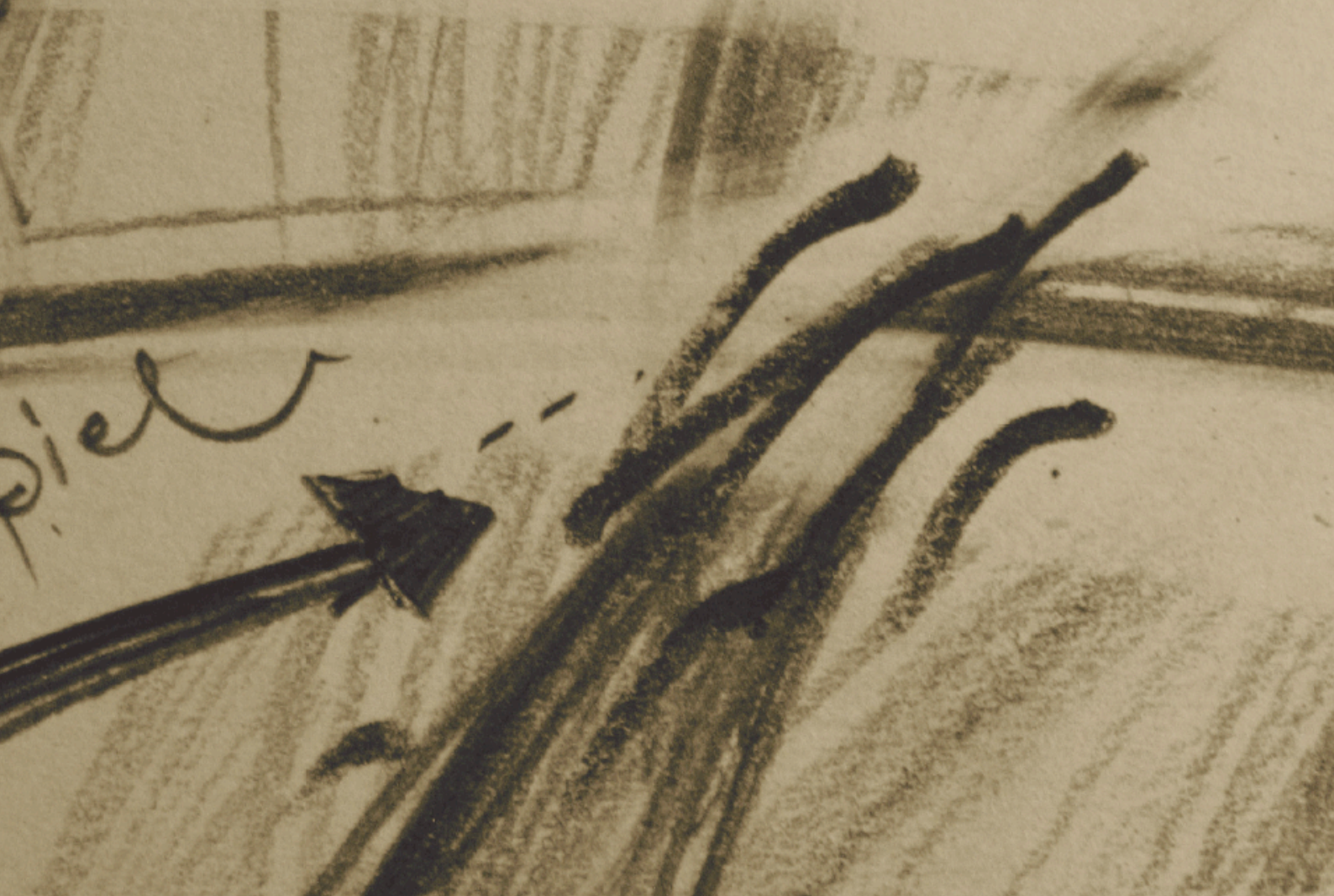
Moitas grazas!

Abreviaturas

- ADN: *Ácido Desoxirribonucleico*
- AIMs: *Ancestral Informative Markers*
- ASIP: *Agouti Signaling Protein*
- AUC: *Area Under the Curve*
- BGA: *BioGeographical Ancestry*
- dNTP: *Deoxyribonucleotide Triphosphate*
- EMSA: *Electrophoretic Mobility Shift Assay*
- EP: *Exclusion Power*
- EVC: *External Visible Characteristics*
- GWAS: *Genome Wide Association Study*
- HGDP: *Human Genome Diversity Project*
- HLTf: *Helicase Like Transcription Factor*
- LD: *Linkage Disequilibrium*
- LEF1: *Lymphoid Enhancer Binding Factor 1*
- LOD: *Logarithm (base 10) of odds*
- LSBL: *Locus Specific Branch Length*
- MALDI-TOF: *Matrix-Assisted Laser Desorption/Ionization Time of Flight*
- MDR: *Multifactor Dimensionality Reduction*
- MITF: *Microphthalmia-Associated Transcription Factor*
- MSH: *Melanocyte Stimulating Hormone*
- OMIM: *Online mendelian inheritance in man*
- ORF: *Open Reading Frame*
- PCA: *Principal component analysis*
- PCR: *Polymerase Chain Reaction*
- POMC: *Proopiomelanocortin*
- RFLP: *Restriction fragment length polymorphism*
- RMP: *Random Match Probability*
- SNP: *Single Nucleotide Polymorphism*
- SLPs: *Single Locus Probes*
- STR: *Short Tandem Repeat*
- SWI-SNF: *SWitch/Sucrose non Fermentable*
- TMRCA: *The Most Recent Common Ancestor*



1. INTRODUCCIÓN



I. Evolución de la genética forense y sus herramientas

La Medicina Legal (o Ciencias Forenses), representa una especialidad que tiene como objeto ayudar a resolver asuntos legales no sólo relacionados con procesos penales sino también con casos civiles. A pesar de tratarse de una ciencia multidisciplinaria, ha sido una de sus herramientas en particular quien ha revolucionado la investigación forense en los últimos años: El análisis del ADN. El objeto de los genetistas forenses es identificar con la mayor certeza posible el origen de una muestra biológica (Jobling-Gill, 2004), a partir del análisis de los polimorfismos presentes en esta molécula.

La evolución en el campo de la genética forense comenzó con el descubrimiento de las variantes estudiadas en los grupos de sangre AB0 (Landsteiner, 1990), y su posterior aplicación para el estudio de filiaciones así como en casos criminales, hasta la revolución de la huella de ADN en 1984, con el descubrimiento de *loci* hipervariables conocidos como minisatélites, (Jeffreys et al., 1985) generando los patrones multibanda que hoy día conocemos como huella génica. El uso de la huella de ADN ha sido empleada en pruebas de filiaciones, así como en casos criminales con el análisis de SLPs (*Single Locus Probes*). Posteriormente, tras el descubrimiento de los **STRs** (*Short Tandem Repeats*) o secuencias repetidas en tándem, junto con las tecnologías de secuenciación automatizadas, se han ido consolidando los sistemas para la identificación de individuos, teniendo como principal ventaja el gran poder de discriminación debido a su elevado grado de polimorfismo y por consiguiente informatividad, así como la sensibilidad en la detección de las secuencias de interés (Jobling-Gill, 2004). El análisis estadístico de dichos sistemas está basado en el establecimiento de una probabilidad de identidad entre perfiles de ADN, y dicho análisis se encuentra implementado hoy día en casuística forense para pruebas de identificación, filiación y criminalística (Evett-Weir, 1998).

Debido a las características anteriormente mencionadas, así como a la eficacia y al crecimiento de las bases de datos de STRs en criminalística, las cuales contienen en algunos países hasta millones de perfiles empleados en rutina, es poco probable que otra clase de marcadores genéticos tengan un mayor impacto en la comunidad

forense. Otros polimorfismos como los SNPs (*Single Nucleotide Polymorphisms*) y otros sistemas de genotipado de ADN deben por tanto, ser considerados como suplementos y no sustitutos de los STRs, especialmente en aquellos casos de identificación y filiación de gran complejidad (Butler, 2009; Phillips et al., 2008), así como también para el análisis de ADN degradado (Fondevila et al., 2008).

Sin embargo, cuando no se encuentra disponible un perfil genético de referencia con el que sea posible realizar comparaciones - no existe un sospechoso, o no existe correspondencia con ningún perfil en las bases de datos disponibles - estos sistemas presentan limitaciones, y es allí cuando cobran gran importancia otros marcadores genéticos que puedan aportar información adicional que permita reconstruir, al menos en parte, la apariencia física del individuo en cuestión, de tal forma que sea posible reducir el universo de sospechosos presentes en la población.

En los últimos años, se han estudiado estos marcadores genéticos alternativos que son ya de utilidad para un cierto número de casos, en los que los marcadores convencionales tal y como se ha mencionado anteriormente, presentan una utilidad limitada. Esto, ha permitido una variedad de aplicaciones al estudio del ADN humano con fines forenses. Una de estas aplicaciones, y el tema de estudio en el presente trabajo, es la predicción de características visibles externas de un individuo o **EVCs** (*External Visible Characteristics*). Así mismo, en este trabajo será abordado el estudio de la inferencia de grupos ancestrales (Shriver-Kittles, 2004), el cual igualmente representa una fuente de información de interés forense, a través del estudio de **AIMs** (*Ancestral Informative Markers*), explorando tanto los SNPs autosómicos, los cuales permiten la inferencia del grupo ancestral biogeográfico o BGA, así como los marcadores de linajes por análisis de ADN mitocondrial.

En resumen, existen herramientas alternativas empleadas en la genética forense que actualmente permiten la obtención de información adicional y/o complementaria a la obtenida mediante el estudio de STRs, entre las que figuran los polimorfismos de una base o SNPs, marcadores genéticos que serán abordados en el presente estudio. En el apéndice 1 se adjunta una revisión sobre estos y otros marcadores alternativos.

1.1. Single Nucleotide Polymorphisms (SNPs)

Los polimorfismos de una sola base o SNPs representan la mínima alteración en longitud que puede experimentar la secuencia de ADN de un individuo, y a su vez la alteración con mayor frecuencia de aparición en el genoma humano (Taillon-Miller et al., 1999; Spalvieri-Rotenberg, 2004). De acuerdo a los datos obtenidos del Proyecto Genoma Humano, existe aproximadamente un SNP por cada 100-300 pares de bases (U.S Program, 2008) y actualmente se han descrito millones de estos marcadores a lo largo del genoma (<http://hapmap.ncbi.nlm.nih.gov>). La alteración en una base nucleotídica puede dar origen a transiciones que corresponden al cambio que ocurre de una purina a otra purina (Ej. A por G) o de una pirimidina a otra, (Ej. C por T), y las transversiones que representan el cambio que ocurre de una purina a una pirimidina o viceversa (Ej. A por T, o G por C). También son llamados SNPs a las inserciones y deleciones en una sola base conocidas como *Indels* (aunque este término también puede abarcar un mayor número de bases). Estos cambios nucleotídicos deben ocurrir al menos con 1% de incidencia en la población para ser considerados como polimorfismos.

Dos procesos fundamentales pueden causar las mutaciones por sustitución de una base: la incorporación errónea de nucleótidos durante la replicación, y la mutagénesis causada por modificaciones químicas de bases o daño físico debido a por ejemplo, radiación o ionización ultravioleta (Jobling et al., 2004).

Las sustituciones de una base pueden tener un amplio rango de efectos sobre la expresión y regulación de genes (cuando se encuentran fuera del marco abierto de lectura) así como en la estructura y función de las proteínas (cuando están contenidas dentro del marco abierto de lectura). Sin embargo, la redundancia del código genético ofrece una amortización frente a los posibles efectos deletéreos, ya que muchos aminoácidos están codificados por más de un codón y en general, la tercera posición del codón es insensible ante sustituciones de una base. Cuando las sustituciones de una base se encuentran contenidas específicamente dentro de un exón se conocen como SNPs codificantes (cSNPs), los cuales en la mayoría de los casos ocasionan proteínas con expresión, estructura, o funciones biológicas alteradas. Las sustituciones que no producen alteraciones en los aminoácidos son conocidos

como sitios silentes o sustituciones sinónimas, y cuando los SNPs dan lugar a cambios aminoacídicos estos son llamados *missense*, o *nonsense* cuando la sustitución origina un codón de terminación (Jobling et al., 2004). Debido a que sólo de 3 a 5% del ADN corresponde a secuencias codificantes, la mayoría de los SNPs se encuentran fuera de estas regiones. Sin embargo, aun en ausencia de significado funcional, su proximidad a un determinado gen alterado, con el que segrega en forma conjunta los convierte en indicadores útiles para detectar por ejemplo, potenciales anomalías génicas (Spalvieri-Rotenberg, 2004). Su localización en zonas definidas del ADN permite elaborar mapas cromosómicos indicadores de su posición relativa respecto a genes conocidos, y a la vez, caracterizar las interacciones con otros genes.

Se conoce que las frecuencias alélicas de estos marcadores varían en gran medida entre poblaciones, como dentro de ellas. Esta disparidad en la frecuencias ocurre debido a que cada población posee su propia historia genética, así como diversos patrones de migración geográfica, de expansión y por las variaciones estocásticas que contribuyen igualmente a que existan estas diferencias (Cardon-Palmer, 2003). Por otra parte, la densidad de SNPs puede variar de acuerdo a la región cromosómica (Venter et al., 2001). La historia genealógica de un *locus*, puede afectar no sólo a las diferencias de frecuencias, sino también a la densidad de SNPs, siendo el determinante primario de las variaciones en la diversidad de su secuencia, incluso en mayor medida que su tasa de mutación. (Jobling et al., 2004; Reich et al., 2002).

I.1.a. Tecnologías para la detección y genotipado de SNPs

Existen diferentes estrategias que pueden ser aplicadas al descubrimiento de nuevos SNPs. El método más directo es la secuenciación, seguido por la amplificación de regiones *locus*-específica en el genoma. Actualmente, hay un gran número de tecnologías que han sido desarrolladas para automatizar el genotipado de SNPs, las cuales varían en parte de acuerdo a su nivel de rendimiento (relación entre el número de marcadores y número de muestras que sea requerido analizar), por ejemplo: *Taqman* (1-5 SNPs), *SNapShot* (1-20 SNPs), *Sequenom* (1-150 SNPs), *SNPlex* (48-400 SNPs), y *Affymetrix* que junto con Illumina (5000-900000 SNPs) continúan en

expansión, pudiendo tener una cobertura del orden de millones de SNPs. Igualmente, estas técnicas de genotipado pueden ser clasificadas en función del mecanismo molecular que utilicen dentro de uno de los siguientes grupos: hibridación alelo-específica, extensión del *primer*, ligamiento de oligonucleótidos, y corte invasivo (Sobrino-Carracedo, 2005).

Entre los métodos comúnmente empleados en el genotipado SNPs en la casuística forense están la secuenciación, y la minisequenciación por *SNaPShot*. Una de las ventajas que presentan estos métodos es que las muestras pueden ser analizadas en secuenciadores automatizados como los ABI 310 o 3100, los cuales se encuentran disponibles en la gran mayoría de los laboratorios forenses, ya que dichos instrumentos también son empleados para el análisis de rutina de STRs. Sin embargo, cuando se requiere el análisis genético de poblaciones, por ejemplo para implementar bases de datos, el empleo de las tecnologías de alto rendimiento facilitan la automatización y el análisis de estos marcadores en un gran número de muestras (Sobrino-Carracedo, 2005).

- Secuenciación: En sus inicios, la secuenciación de ADN fue realizada mediante el empleo de métodos químicos que producían una modificación base-específica, seguido por un corte en la secuencia de ADN. En la actualidad, los métodos enzimáticos de primera generación son los más utilizados en el contexto forense, siendo el más conocido el método de secuenciación Sanger (Sanger et al., 1977). El ADN provisto en forma de cadenas sencillas es empleado por la enzima ADN polimerasa para así sintetizar una nueva hebra complementaria a la secuencia de ADN original. Durante la reacción de secuenciación se emplean cebadores o *primers* universales, nucleótidos (dNTPs) marcados (generalmente por fluorescencia) y dideoxinucleótidos trifosfatos (ddNTPs), los cuales son análogos a los dNTPs pero difieren en que carecen de un grupo hidroxilo en el carbono 3' así como en el carbono 2'. Estos ddNTPs son incorporados en la hebra de ADN que está en construcción formando enlaces fosfodiéster entre su carbono 5' y el carbono 3' del nucleótido previamente incorporado. Sin embargo, al carecer del grupo hidroxilo en el carbono 3', los ddNTPs que son incorporados no pueden establecer enlaces fosfodiéster a continuación

causando una terminación en la síntesis de la cadena. Dicha terminación, ocurre al azar en una de las muchas posiciones que contiene una de las cuatro bases en cuestión. Por lo tanto, cada reacción es en realidad parcial, en la cual la terminación de la cadena ocurre al azar en una de las posibles bases en cualquiera de las hebras de ADN. Como resultado, cada una de las cuatro reacciones base-específica generará una colección de fragmentos de ADN marcados de diferentes tamaños con un mismo extremo 5' pero con diferentes extremos 3'. Una mejora en la visualización de los fragmentos generados ha sido el desarrollo de procedimientos automatizados para la secuenciación con fluorescencia (Wilson et al., 1990), en la cual los ddNTPs poseen fluoróforos. En estos casos, durante la electroforesis un sensor detecta y graba las señales fluorescentes que se van generando; el *output* obtenido muestra perfiles de intensidad para cada uno de los fluoróforos (Strachen-Read, 1996). En la actualidad, existen *kits* comerciales disponibles para realizar la reacción de secuenciación a partir de un producto de PCR específico de una o varias regiones de interés. Previo a la reacción de secuencia, dicho producto de PCR requiere de un paso de purificación así como luego de la PCR. La purificación puede ser de tipo física (mediante el empleo de membranas adaptadas a placas) o enzimática (con el uso del *Exosap*, una combinación de dos enzimas, una de ellas con actividad exonucleasa que degrada las cadenas sencillas de los primers, junto con la fosfatasa alcalina SAP (*Shrimp Alkaline Phosphatase*) que destruye los dNTPs que no se unieron en la reacción). Debido a la gran demanda en secuenciación de genomas y a la necesidad de disminuir costos y tiempo, en la actualidad se han implementado tecnologías de secuenciación de segunda y tercera generación conocidas como ultra-secuenciación o *next-generation sequencing*, que analizan secuencias obtenidas rápidamente a través de pequeños fragmentos de ADN, dicha característica ha permitido obtener información incluso de material genético antiguo, el cual normalmente se encuentra muy fragmentado para ser analizado con técnicas de secuenciación convencional. Un ejemplo del empleo de esta tecnología con ADN antiguo ocurrió en el año 2008, donde se obtuvo la secuencia completa del genoma mitocondrial de un Neanderthal (Green et al., 2008). Algunas de las limitaciones de la técnicas de ultra-secuenciación, son los requerimientos

de una nueva instrumentación, así como la complejidad en el manejo bioinformático de los datos.

- Minisecuenciación: En general, la minisecuenciación está basada en la precisión en la incorporación de nucleótidos por parte de la DNA polimerasa, mediante la unión de un *primer* a la secuencia diana de ADN inmediatamente adyacente al SNP que es extendido por la polimerasa (Fig. 1a). La capacidad de generar reacciones *multiplex* depende de la tecnología utilizada, entre las que están la espectrofotometría de masas (MALDI-TOF), y la más comúnmente empleada en la comunidad forense: la detección por fluorescencia con *SNapShot*. La minisecuenciación por *SNapShot*, implica la extensión de un *primer* alelo-específico empleando ddNTPs marcados fluorescentemente. Se utilizan cuatro fluorocromos diferentes de tal forma que se correspondan con las cuatro bases nucleotídicas posibles que pueden ser incorporadas durante la extensión. Existen tres pasos primordiales en la realización de una minisecuenciación: amplificación, extensión del *primer* alelo específico y análisis en un secuenciador automático de electroforesis capilar (Tully et al., 1996). Primero, la región flanqueante de cada SNP es amplificada mediante PCR, dicha amplificación puede ocurrir en reacciones múltiplex lo cual permite analizar simultáneamente varios marcadores de este tipo. Los dNTPs y *primers* remanentes posteriores a la PCR son eliminados por la adición de *Exosap*. Este paso es necesario para evitar futuras interacciones con las próximas reacciones de extensión. En el segundo paso, se agrega al producto de PCR purificado con *Exosap* una mezcla que contiene: los *primers* de extensión, los ddNTPs marcados y una polimerasa. El *kit* de *SNapShot* de *Applied Biosystems* (AP) contiene los ddNTPs marcados fluorescentemente, el tampón o *buffer* y la polimerasa, permitiendo que cada usuario pueda implementar a esta mezcla su propio diseño de *primers*. Los *primers* de extensión están diseñados para anclar sobre las regiones adyacentes a cada SNP de interés, de tal manera que sea posible la incorporación de un ddNTP que corresponderá a la base nucleotídica que está siendo interrogada. A continuación de la reacción de extensión, los productos son tratados con la enzima SAP para remover los ddNTPs marcados que no fueron incorporados. Los fluorocromos empleados

durante la reacción y que posteriormente serán de utilidad para distinguir las bases nucleotídicas detectadas son: dR6G de color verde, dTAMRA™ de color amarillo aunque en el *software* es detectado como color negro (para generar mayor contraste), dR110 de color azul, y el DRox™ de color rojo. De tal manera que la presencia de un pico azul en el electroferograma indicará que una G (ddGTP) ha sido incorporada por la polimerasa en el SNP correspondiente. De igual manera son analizadas las otras bases en el electroferograma con sus respectivos fluorocromos (A para el verde, C para negro, y T para el rojo). Adicionalmente, es posible analizar en las plataformas electroforéticas un marcador interno (visualizado con un fluorocromo naranja) para corregir las migraciones en las corridas. Como se ha mencionado anteriormente, es posible analizar múltiples *primers* simultáneamente por la adición de secuencias nucleotídicas de diferentes longitudes al extremo 5' de éste, de tal forma que cada *primer* extendido tenga una movilidad diferente por el tamaño de la cola de bases añadidas, cuya secuencia no deberá hibridar con ninguna región nucleotídica que se encuentre presente en la reacción. Para verificar la posibilidad de hibridación así como la formación de horquillas o *hairpins* de estas secuencias existen programas que permiten hacer simulaciones *in silico* empleadas durante el diseño de estas secuencias, tal es caso del programa gratuito *Auto Dimer* (Vallone-Butler, 2004). El análisis de los electroferogramas es realizado con programas como *Genotyper* o *GeneMapper* (AP). Entre las limitaciones de esta técnica está la posible presencia de artefactos producto de una purificación deficiente durante la adición de *Exosap* o SAP, las cuales pueden interferir con las reacciones de extensión (Butler, 2011) ocasionando *dye blobs*.

- Tecnologías de alto rendimiento: Dentro de las tecnologías de alto rendimiento, los *arrays* de alta densidad permiten que cientos, miles, e incluso millones de SNPs sean analizados simultáneamente por muestra. En el campo forense, estas tecnologías presentan como principal limitación, el requerimiento mínimo de concentración de ADN que generalmente no suele ser alcanzado a partir de un vestigio biológico encontrado en una escena del crimen. Estos *arrays* están siendo empleados en la investigación forense principalmente para realizar estudios de asociación en genomas completos o

GWAS , relacionando por ejemplo variantes génicas a rasgos físicos como la pigmentación humana (Han et al., 2008; Kayser et al., 2008; Stokowski et al., 2007; Sulem et al., 2007; Sturm et al., 2008), lo que resulta de utilidad en la identificación de SNPs que puedan contribuir a la inferencia de características físicas. De igual forma, se han implementado plataformas de genotipado específicas de ciertos grupos de poblaciones (Hoffmann et al., 2011), así como para la inferencia del ancestro biogeográfico (<https://www.23andme.com/>). Estas tecnologías permiten entonces generar bases de datos de interés en la investigación criminal, que a su vez requiere los datos obtenidos a partir de la genética de poblaciones. Entre las plataformas de alta densidad más comúnmente empleadas en la investigación forense se encuentran *Affymetrix* e *Illumina*. *Affymetrix* posee *arrays* de genoma completo siendo el 6.0 el de mayor cobertura con la tecnología *GeneChip*, donde es analizada una muestra por arreglo. Este *chip* consta de *microarrays* que poseen un soporte de cristal el cual contiene millones de sondas distribuidas estratégicamente, permitiendo detectar la presencia o ausencia de secuencias nucleotídicas (de ADN o ARN) en una muestra biológica mediante el empleo de un *scanner*. Estas sondas oligonucleotídicas son sintetizadas por fotolitografía mediante la incorporación de nucleótidos a grupos protectores fotosensibles. El principio por el cual son detectados los cambios en las secuencias nucleotídicas es mediante la hibridación de la sonda a la hebra de ADN, es decir una hibridación alelo-específica (Fig. 1b). Previo a la hibridación, la muestra requiere un procesamiento en el cual es sometida a la acción de dos enzimas de restricción (*Nsp I* y *Sty I*). Posterior a la digestión enzimática, se unen los fragmentos generados con adaptadores complementarios de *primers* de PCR, y luego de la amplificación, los productos sufren una nueva fragmentación por acción de una *DNase*, así como un marcaje con biotina. Esta molécula puede ser detectada por tinción fluorescente, de tal forma que la fluorescencia es emitida cuando ocurre la hibridación. En cuanto a la selección de los marcadores empleados en estos *arrays*, esta plataforma puede utilizar *Tag* SNPs (del proyecto HapMap), SNPs distribuidos al azar, o la combinación de ambos criterios. De igual manera, es posible incorporar a un arreglo SNPs de tipo funcionales así como específicos de una población. Reciente ha sido implementado el *Affymetrix*

Axiom, en donde por cada chip es posible analizar hasta 96 muestras empleando una cobertura y tecnología de detección similar al *GeneChip*, además de poseer un contenido actualizado sobre los SNPs que han sido incorporados en sus *arrays*. Por otra parte, la plataforma *Illumina*, emplea la tecnología de extensión alelo-específica (Fig. 1c-d) junto con una ligamiento del oligonucleótido, el cual se basa en una reacción directa de discriminación alélica sobre el ADN genómico y en la especificidad de las enzimas ligasas, que permiten la unión de productos amplificados simultáneamente por *primers* de PCR universales. En la reacción de genotipado, los productos de ligamiento contienen secuencias dirigidas a hibridar con sondas que se encuentran en el *array*. Los productos son marcados fluorescentemente empleando un color diferente para cada alelo del SNP. A continuación de la hibridación al *array*, se emplean los ratios entre las dos señales alelo-específicas emitidas para determinar el genotipo de cada SNP. La detección esta basada en la tecnología *BeadArray*TM donde cada unidad o *bead* representa un elemento sensor de una secuencia particular de ADN, permitiendo analizar más de 1000 SNPs por ensayo (Sobrino-Carracedo, 2005). Otras de las aplicaciones forenses donde se ha empleado estas plataformas, es el análisis de mezclas en criminalística, donde mediante el empleo de estas tecnologías ha sido posible identificar y determinar el número de contribuyentes a una mezcla compleja de ADN (Homer et al., 2008). Recientemente, la plataforma de genotipado a larga escala *Affymetrix* 6.0, ha sido también empleada para el análisis de filiaciones entre parientes lejanos, particularmente en un caso donde participaron primos segundos, el cual no sería posible resolver empleando los marcadores comúnmente usados en la rutina forense (Lareu et al., 2012).

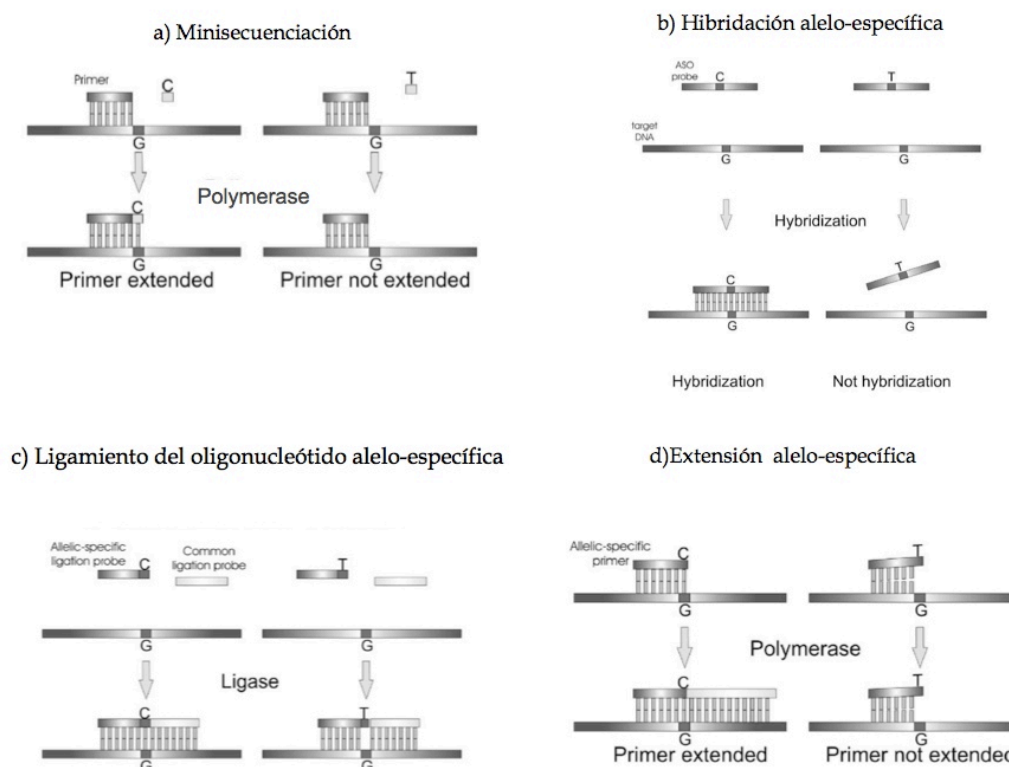


Fig.1. Principales mecanismos moleculares empleados en el genotipado de SNPs con aplicación forense: Extensión del *primer*, que incluye tanto la Minisequenciación (a) fundamento del *SNapShot*, como la extensión alelo-específica (d) que utiliza *Illumina* junto con la ligamiento de oligonucleótidos alelo-específica (c). El mecanismo molecular de genotipado en *Affymetrix* está basado en una hibridación alelo-específica (b). Basado en Sobrino y col. 2005.

I.1.b. Bases de datos de SNPs

En la actualidad se encuentra disponible un amplio listado de SNPs validados en diversas bases de datos *online* como dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>), además de en catálogos como *SPSmart* (<http://spsmart.cesga.es/>) (Amigo et al., 2008). Esta última herramienta posee la ventaja de incluir diversas bases de datos como *1000 Genomes*, *HapMap*, *Perlegen*, *CEPH-HGDP* (de la Universidad de Stanford, y Michigan). De igual manera, mediante el uso de este catálogo, es posible obtener información sobre las frecuencias alélicas que presentan estos marcadores en diversas poblaciones del mundo. El criterio en la selección de SNPs en bases de datos *online* como las presentadas en la tabla 1 para el desarrollo de

ensayos forenses, requiere que los marcadores empleados posean ciertas características que le confieran especificidad en el campo requerido.

Tabla 1. Principales bases de datos de SNPs. NCBI *National Center for Biotechnology Information*, NHLBI *National Heart Lung and Blood Institute*, PGA *Program for Genomic Applications*. Phillips , 2009 (Phillips, 2009).

Database	Host organization	Gateway URL for initiating SNP data searches
dbSNP	NCBI	http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp
HapMap	The HapMap Consortium	http://www.hapmap.org/cgi-perl/gbrowse/
Ensembl	EMBL-EBI/Sanger Center	http://www.ensembl.org/Homo_sapiens/index.html
Santa Cruz	University of California, Santa Cruz	http://genome.ucsc.edu/cgi-bin/hgGateway
Perlegen	Perlegen Sciences	http://genome.perlegen.com/browser/index_v2.html
Assays-on-Demand	Applera (Applied Biosystems)	https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=ABGTKeywordSearch&catID=600769
SeattleSNPs	US NHLBI (PGA)	http://gvs.gs.washington.edu/GVS/

Para el diseño de un ensayo de identificación individual como por ejemplo, el 52-plex (Sanchez et al., 2006) es recomendable seleccionar al menos 50 SNPs, tomando además en consideración factores como el ligamiento (ya que se requieren varios SNPs por cromosoma). Igualmente, deberá analizarse la calidad de las secuencias adyacentes, así como el grado de polimorfismo de los SNPs de interés para el diseño de un *multiplex* (Phillips, 2005). Estos criterios, junto con otros de mayor especificidad, pueden ser seleccionados en las diversas bases de datos *online* presentados en la tabla 1, así como en el catálogo *SPSmart* de interés en la comunidad forense.

I.1.c. Aplicaciones forenses de los SNPs

Los SNPs han sido considerados como potenciales marcadores genéticos con fines forenses por varias razones, una de ellas debido a que los productos de PCR generados pueden tener una longitud promedio menor a 100 pb, lo que significa que estos marcadores permiten obtener información a partir de material genético degradado en comparación con los STRs e incluso de algunos miniSTRs, cuyos

productos de amplificación podrían oscilar entre 100-400 y 80-200 pb respectivamente (Butler et al., 2002). Estos polimorfismos de una sola base al tratarse de marcadores en su mayoría bialélicos, pueden ser analizados en reacciones *multiplexes*, que incluyen un mayor número de marcadores por reacción en comparación con los STRs, lo cual también facilita su automatización en las diversas plataformas de genotipado antes mencionadas así como su interpretación y análisis estadístico. Sin embargo, esta misma ventaja supone a la vez una limitación, ya que al tener generalmente dos alelos, no resultan tan informativos para identificación individual como los STRs comúnmente empleados, por lo que es necesario analizar un mayor número de SNPs (50-100) para así alcanzar el mismo nivel de discriminación (Gill, 2001). Sin embargo, también se han identificado SNPs trialélicos los cuales incrementan parcialmente el nivel de discriminación tanto en la identificación individual como en el estudio de mezclas. Recientemente, los SNPs han sido categorizados de acuerdo a su aplicación en el campo forense en cuatro grupos (Budowle-van Daal, 2008; Butler et al., 2008):

- SNPs informativos de identificación individual: Como se ha mencionado previamente, estos marcadores representan un suplemento a las pruebas de identificación de STRs, especialmente en aquellos casos de parentescos complejos o cuando el material genético disponible se encuentra en estado de degradación o fragmentación. En el año 2003, se llevó a cabo el proyecto *SNPforID* con el objeto de desarrollar un múltiplex de 52 SNPs loci (Sanchez et al., 2006) el cual ha sido ampliamente estudiado en diversas poblaciones del mundo (Drobnic et al., 2010), así como validado para aplicaciones forenses (Borsting et al., 2009) y para pruebas de paternidad (Borsting et al., 2008).
- SNPs informativos de linaje: Los SNPs de linaje permiten definir haplogrupos definidos en bloques haplotípicos tanto en el cromosoma Y como en el ADN mitocondrial, y son considerados en cada caso como un único marcador de linaje, los cuales poseen diversos haplotipos en comparación con los alelos de un *locus*, lo que puede ser también un complemento de valor en las pruebas de parentesco. Para el caso de cromosoma Y así como en el X, los SNPs son de gran utilidad debido a su

baja tasa de mutación en comparación con los STRs (Brion et al., 2005; Alvarez-Iglesias et al., 2007).

- SNPs informativos de grupo ancestral: Aunque para la identificación individual se requiere un mayor número de SNPs para poder alcanzar niveles de discriminación similares a los obtenidos con STRs, para el caso de la inferencia de grupos ancestrales, son los SNPs quienes poseen un gran potencial en comparación con el resto de los marcadores genéticos. Por ejemplo, estos AIMs presentan mayor poder en la diferenciación de grupos de poblaciones (Phillips et al., 2007) en relación con los microsatélites (Rosenberg et al., 2002). Aunque existen estudios sobre el empleo de perfiles de STRs autosómicos para diferenciar grupos continentales, se ha recomendado acompañar estos análisis con el estudio de otros AIMs como los SNPs autosómicos, los cuales aportan mayor fiabilidad al análisis (Phillips et al., 2011), ya que las inferencias de poblaciones realizadas con STRs pueden variar dependiendo del grupo ancestral con el que sea comparado (Graydon et al., 2009). Las variaciones en el ADNmt y en cromosoma Y, están limitados a dar información de tipo filo-geográfica. Por otra parte los indels, aunque ofrecen ventajas en cuanto a la simplicidad en el genotipado frente a los SNPs y en sus productos de amplificación de menor tamaño frente a los STRs, aún requieren ser analizados en un mayor número de grupos de poblaciones para comprender y evaluar el valor de la estimación de grupos ancestrales (Santos et al., 2010). Los SNPs autosómicos, debido a su estabilidad, densa distribución y amplio rango de frecuencias diferenciales en poblaciones del mundo, aportan información complementaria a los marcadores de linajes en el análisis de la historia evolutiva de las poblaciones, y tienen la ventaja de ofrecer una sensibilidad de detección superior en muestras forenses (Phillips et al., 2007).
- SNPs informativos de rasgos físicos: Aunque existen una variedad de rasgos físicos que presentan un fuerte componente genético, la característica física más ampliamente estudiada es la pigmentación. Dentro de este rasgo, ha sido posible el estudio de SNPs en el *MC1R* que determinen la predicción del cabello pelirrojo (Grimes et al., 2001), así como la aplicación en la inferencia

del color de ojos azules y marrones determinados principalmente por *HERC2* (Liu et al., 2009; Walsh et al., 2011b). El análisis de SNPs asociados al color de la piel puede ser también una vía útil para detectar AIMs, teniendo especial cuidado con aquellos SNPs sujetos a un evento de selección (Salas et al., 2006), ejemplo de ello son las variantes en el gen *SLC24A5* que han sido asociadas a índices de melanina en pieles claras, y sin embargo, este hallazgo también se ha correlacionado con el estudio del grupo ancestral Europeo. En la tabla 2 se resumen las principales características y aplicaciones que tienen los SNPs frente a otros marcadores, en donde puede observarse que exceptuando en la inferencia del sexo -- se destaca la exclusividad que poseen los SNPs en la inferencia de rasgos físicos humanos en comparación con otros marcadores genéticos.

Recientemente, se ha publicado un trabajo sobre el empleo de “SNPs nucleosómicos” con aplicación en el análisis de material genético de alta degradación, basado en las evidencias existentes sobre la función protectora que ejerce el complejo de histonas ubicado en los nucleosomas, ante los procesos de degradación celular, empleando por tanto un conjunto de SNPs localizados en las regiones genómicas que circundan estas estructuras. Tras el empleo de este multiplex, se ha encontrado un ligero incremento en el éxito del genotipado de muestras forenses frente a otros sistemas de identificación individual como el 52-plex y los miniSTRs, representando entonces una nueva alternativa en este campo de investigación (Freire-Aradas et al., 2012).

Tabla 2. Comparación de los SNPs con otros marcadores genéticos. Basado en Butler, 2011

Característica	STRs	SNPs	Marcadores de Linajes	Indels
Frecuencia en el genoma	~ 1/15 Kb	~ 1/0,3-1 Kb	Cromosoma Y: 1/ genoma. ADNmt >1000/1 DNA nuclear	~ 15/100 Kb
Grado de información individual	Alta (~ 15 marcadores)	Baja, 20-30% de los STRs. (~ 52 marcadores)	No tiene. Sólo linajes patri-lineal(Y) y matri-lineal (ADNmt)	Media (~ 38 marcadores)
Tasa de mutación	~ 1/10 ³	~ 1/10 ⁸	~1/10 ³ (Y-STRs) ~1/10 ⁹ (Y-SNPs) ADNmt 5-10veces>ADNnuclear	~ 1/2,3 x10 ⁻⁹ (indels cortos)
Tipo de marcador	Di, tri, tetra, penta-repeticiones de nucleotidos.	Bialélicos: A/G, C/T, A/T, C/G, T/G, A/C	Haplotípicos.	Inserciones o deleciones, 71% de 1, 2, 3, 4 nucleotidos
Número de alelos/haplotipos	Usualmente entre 5 y 20	Generalmente 2, aunque existen algunos tri y tetra-alélicos	Variable.	Generalmente bialélicos, llamados “cortos” y “largos”
Métodos de detección en forense	PCR con <i>primers</i> marcados, gel, electroforesis capilar.	Secuenciación, Tec. alto rendimiento, electroforesis capilar.	Variable, dependiendo del tipo de marcador empleado.	PCR con <i>primers</i> marcados, Secuenciación, electroforesis capilar. Tec. alto rendimiento
Capacidad de reacciones multiplex	> 10 marcadores con coloración fluorescente	~ 1000 SNPs, dependiendo de la tecnología empleada	Variable.	Similar a los SNPs.
Tamaño del producto de PCR	~ 75-400 pb	~ < 100 pb	Variable, dependiendo del tipo de marcador empleado.	~ <160 pb
Predicción de grupos ancestrales	Limitado	SNPs autosómicos asociados a BGA	Sólo aportan información filo-geográfica.	Se ha evaluado el BGA en ciertas poblaciones
Predicción de características externas visibles (EVCs)	No	Existe asociación de SNPs a ciertos rasgos físicos	No	No

II. La predicción de rasgos físicos en un contexto forense

El retrato hablado bajo un contexto forense, reúne la información que pueda ser empleada para el conocimiento sobre la apariencia física de un individuo en cuestión, y resulta de gran utilidad en la investigación criminal, ya que representa la principal descripción ofrecida por un testigo ocular (Heaton-Armstrong, 1995). Una de las principales razones por la que muchos científicos forenses tienen interés en estudiar el ancestro biogeográfico de un individuo, es justamente para poder reconstruir su apariencia física. Sin embargo, la información contenida detrás del análisis de la apariencia física puede o no estar definida por el grupo ancestral. Ciertas características físicas se encuentran ampliamente distribuidas a nivel mundial, como los niveles de pigmentación en la piel (Frost, 2007), cabellos y ojos (Frost, 2006), e incluso rasgos faciales como la forma de los ojos y nariz, y otras características como tamaño de los labios, la habilidad de las manos, la orientación en la que crece el cabello, la forma del lóbulo de la oreja, etc. (Pulker et al., 2007). La variabilidad en los rasgos físicos puede encontrarse también entre individuos que poseen una proporción de componentes ancestrales múltiples y en estos casos, aún pudiendo presentar un componente mayoritario, su apariencia física generalmente será diferente a la esperada (Jobling et al., 2004; Valenzuela et al., 2010). Es por ello, que la investigación forense posee también gran interés en el estudio de características físicas, y su predicción a partir del grupo ancestral biogeográfico deberá ser interpretado como una inferencia indirecta, basada en la observación de un carácter físico distribuido como función del grupo ancestral, pero que se fundamenta en el análisis de AIMs y no de *loci* funcionales que conlleven a la expresión de un rasgo físico dado (Frudakis, 2008).

La inferencia de características físicas en un contexto forense a través del análisis de marcadores genéticos presentes en el ADN, ha sido acuñado como *Forensic DNA Phenotyping* (FDP) (Kayser-de Knijff, 2011) o *Forensic Molecular Photofitting* (Frudakis, 2008), y reúne la información obtenida tanto del grupo ancestral biogeográfico o BGA (*BioGeographical Ancestry*), como de las características visibles externas de un individuo o EVCs (*External Visible Characteristics*). El análisis de FDP tiene aplicaciones también en el estudio de restos antiguos y en la identificación de víctimas en desastres naturales. Cabe destacar que el estudio de

FDP puede contribuir a evitar errores en la identificación de sospechosos dado que a consecuencia del trauma, un alto porcentaje de víctimas de un delito pueden equivocarse en las ruedas de reconocimiento, o presentar testimonios confusos (Heaton-Armstrong, 1995), mientras que el valor en la predicción de características físicas con marcadores genéticos poseerá en todo caso un respaldo estadístico (Kayser-de Knijff, 2011). Todo esto, teniendo siempre en cuenta que la información obtenida a partir de un estudio con FDP se referirá a un tipo de análisis diferente al que puede ofrecer una identificación individual, representando un complemento de valor frente a los análisis de rutina forense.

II.1. Marcadores informativos de grupos ancestrales (AIMs)

La definición de población humana puede estar basada por ejemplo, en grupos separados por una barrera geográfica, pero también bajo otros criterios subjetivos que puedan incluir un fenotipo dado, la ecología, la distribución territorial de los individuos, así como por características lingüísticas y culturales, entre otras (Pritchard et al., 2000; Jobling et al., 2004). Aunque los humanos somos aproximadamente un 99,9% idénticos en nuestra secuencia de ADN, existe un 0,1% de esta secuencia que define nuestra individualidad genética. De este porcentaje, tan sólo una pequeña porción del genoma (0,01%), corresponde a las diferencias entre poblaciones.

Por su parte, la definición de **ancestro** puede estar contemplado a nivel genealógico así como genético. El ancestro genealógico depende de las poblaciones de donde provengan nuestros antepasados, y tiene a su vez una estrecha relación con el “nivel esperado de grupo ancestral” determinado por la historia genealógica o *pedigree* de un individuo. El ancestro genealógico de una persona en particular, teóricamente es el promedio de la composición ancestral de sus padres, siendo considerada como una variable esperada. Sin embargo, debido a eventos como la recombinación genética y la segregación independiente, en realidad se observaran diferencias en la medida del “ancestro genómico” de este individuo, así como si éste es comparado por ejemplo, con otros parientes cercanos como sus hermanos. Esta variación sobre el nivel actual del grupo ancestral de un individuo basado en marcadores genéticos, permite realizar análisis estadísticos, debido en parte, a que

dichos marcadores pueden ser específicos de una población o poseer frecuencias alélicas diferenciales entre poblaciones que han sido definidas étnica o geográficamente (Cavalli-Sforza et al., 1994; Dean et al., 1994), además de las características propias de la constitución de cada marcador. Algunos de estos marcadores a nivel fisiológico permiten establecer diferencias entre ciertas poblaciones que pueden ser un reflejo por ejemplo, de una adaptación genética a una condición ecológica, deriva génica o selección sexual.

En general, para medir la estructura de las poblaciones en términos del grupo ancestral, primero es necesario identificar y caracterizar aquellos marcadores cuyas frecuencias alélicas se encuentren ampliamente diferenciadas entre los grupos de poblaciones. Entonces, lo que es medido en un individuo es un valor que podemos llamar ancestro genómico o nivel de ancestro genético. El ancestro genómico es el nivel promedio de mezcla calculado en el genoma, empleando marcadores informativos de grupos ancestrales, y esta medida depende de cuan informativos sean estos marcadores, de cuántos de ellos sean empleados y de cómo se encuentran distribuidos en el genoma (Frudakis, 2008). Estos marcadores fueron definidos inicialmente como “Alelos específicos de población” (PSAs)(Shriver et al., 1997), y actualmente se conocen como **AIMs** (*Ancestral Informative Markers*) (Frudakis et al., 2003b; Frudakis et al., 2003a; Shriver et al., 2003). Estos representan polimorfismos cuyas frecuencias alélicas exhiben una diferenciación entre poblaciones, y pueden ser empleados para inferir el origen geográfico de un individuo.

Diversos estudios han determinado que el empleo de AIMs en la diferenciación de grupos continentales varía dependiendo de la región genómica estudiada. En marcadores autosómicos se ha encontrado una mayor diferenciación dentro de los grupos continentales, mientras que con el ADNmt y cromosoma Y, se ha observado una menor variación dentro de grupos continentales y mayor entre ellos (Jobling et al., 2004).

II.1.a. Estratificación y mezcla de poblaciones

Se conoce que existe **estructura** dentro de una población, cuando las subpoblaciones que la constituyen presentan diferencias entre sus frecuencias alélicas, en

comparación con la población total. Tanto el análisis de ADNmt, los polimorfismos del cromosoma Y, así como los marcadores autosómicos han revelado que existe subestructuración geográfica de las poblaciones humanas (Cavalli-Sforza-Feldman, 2003; Bamshad et al., 2004; Rosenberg et al., 2002; Rosenberg et al., 2005). En el contexto de estudios caso-control, resulta de gran importancia antes de validar una asociación, el poder descartar la presencia de estructura, evaluando si estas variaciones en las frecuencias alélicas se deben a distinciones por la manifestación de un fenotipo dado, y no por diferencias etno-geográficas entre los grupos comparados (Pritchard-Donnelly, 2001).

Por otra parte, las poblaciones no son entidades discretas, puesto que pueden intercambiar sus partes constituyentes - los individuos. Este proceso de **mezcla** (o *admixture*), es principalmente una consecuencia genética común del encuentro entre poblaciones. Las poblaciones vecinas frecuentemente intercambian individuos por un proceso de migración bidireccional. Sin embargo, una tercera población híbrida usualmente no resulta de esta clase de intercambio. El término *admixture* es propiamente empleado cuando ocurre la formación de una población híbrida, por ejemplo, producto de la mezcla de poblaciones ancestrales que han estado en aislamiento previo una de la otra. Por lo tanto, la mezcla de poblaciones puede considerarse mayormente a partir del momento en que estos grupos ancestrales aislados comienzan a estar en contacto (Jobling et al., 2004). El flujo genético entre distintas poblaciones genera patrones de desequilibrio de ligamiento (*admixture linkage disequilibrium*) entre *loci* que tengan frecuencias alélicas diferentes en las poblaciones fundadoras. En aquellas poblaciones de flujo genético continuo, es posible detectar con mayor poder el desequilibrio de ligamiento respecto a grupos donde la mezcla ha ocurrido en una única generación (Pfaff et al., 2001).

En grupos que descienden de la mezcla de dos o más poblaciones aisladas, los cromosomas están formados por segmentos de cada una de las poblaciones parentales, en proporciones relacionadas con las contribuciones relativas de cada grupo originario. La longitud de estos bloques parentales depende del número de generaciones desde que ocurrió la mezcla. Es decir, cuanto más tiempo haya transcurrido desde el momento de mezclarse, más cortos serán los bloques que comparten con sus ancestros debido a la acción de la recombinación. Un ejemplo de

esto lo representan los estudios de *Admixture Mapping* o “mapas de mestizaje”, los cuales consisten en la identificación de regiones cromosómicas que muestran un evidente componente ancestral de la población parental, en aquellos individuos que presentan un fenotipo determinado, mediante la asociación entre el fenotipo y el grupo ancestral del *locus*, empleando estudios comparativos ya sea de sólo casos afectos evaluando las proporciones del grupo ancestral esperado y observado, o por casos y controles. La aplicación de mapas de mestizaje requiere del uso de un panel de AIMs seleccionados a lo largo del genoma, para inferir la proporción de grupos ancestrales de las regiones cromosómicas de los individuos que presenten patrones de mezcla. Se recomienda, ante cualquier estudio de mapas de mestizaje, caracterizar en primer lugar muestras de individuos no mezclados, que representen las poblaciones parentales implicadas en el proceso de mezcla de la población en estudio (Sanchez-Diz-Ramos-Luis, 2010).

Existe una situación de mezcla entre poblaciones donde en lugar de ser considerados todos los *loci* de manera neutral respecto a la información sobre las proporciones de mezcla, existe una contribución desigual por parte de individuos de diferentes sexos y poblaciones ancestrales dando origen a una nueva población con mezcla, evento conocido como **mezcla sesgada por el sexo**. En este caso, hombres y mujeres contribuyen a la mezcla en cantidades desproporcionadas. Un claro ejemplo de este fenómeno son las poblaciones actuales de América, analizadas en el presente estudio:

- América del sur: Los principales grupos ancestrales que han contribuido a la moderna diversidad genética en este territorio son los nativos americanos, los europeos colonizadores y los esclavos de África. Las contribuciones a la mezcla de la población híbrida ocurrieron en tiempos diferentes así como de grupos de poblaciones de distintos tamaños. Si consideramos el caso concreto de **Brasil**, tras la llegada de los portugueses en el año 1500, la población nativo-americana tenía aproximadamente 2,4 millones de habitantes, aunque de acuerdo a datos de FUNAI se estima que esta población pudo haber alcanzado hasta los 10 millones de habitantes para esa época (www.funai.gov.br). Posteriormente, en 1808 llegaron cerca de medio millón de colonos portugueses, predominantemente hombres. En los

siguientes doscientos años, 6 millones de colonos llegaron de los cuales 70% vinieron de Portugal e Italia. Entre mediados del siglo XVI y XVII aproximadamente 4 millones de esclavos Africanos fueron llevados al continente. Más reciente aunque en menor medida, otras poblaciones de Europa, así como grupos provenientes de Siria, Líbano y Japón también representaron flujos migratorios importantes (Jobling et al., 2004). En el estudio de Alves-Silva en el 2002 en población local de Brasil se observó un 97% de linajes masculinos europeos que concuerda con la historia demográfica de esta población (Alves-Silva et al., 2000). En el caso de **Venezuela**, ocurrió un primer mestizaje en el siglo XVI (nativos, europeos y africanos) y un segundo mestizaje más reciente durante 1936-1976 (la población del primer mestizaje y población europea junto con otros grupos minoritarios). En el primer mestizaje, la inmigración española, junto con la posterior introducción de esclavos africanos produjo una disminución en el tamaño de la población nativa de la zona. Se estima que los nativos conformaban 300.000 habitantes a comienzos del siglo XVI, y que en 1994 constituían aproximadamente 60.000 habitantes. En la actualidad, han ido recuperando el tamaño de su población nativa. Diversos historiadores de la época colonial señalan que el primer mestizaje había culminado en el siglo XVIII, es decir que las “castas” no eran tales, por el contrario, se produjo una gran movilidad que origina una heterogénea composición de grupos ancestrales. En el año 1800 cuando la población era de ya 900.000 habitantes, 50% de esta población eran producto de la mezcla de diversos grupos ancestrales, un 20% eran de origen europeo, otro 20% eran nativos, y un 10% eran esclavos africanos. En el segundo mestizaje, una gran inmigración europea, principalmente españoles, portugueses e italianos, fortaleció con su presencia al antiguo mestizaje. La población colombiana, posiblemente más de un millón de inmigrantes, se sumó también a este evento (Morón, 1971). Diversos estudios confirman esta mezcla en distintas poblaciones actuales de Venezuela, tales como en Caracas (Lander et al., 2008) y Maracaibo (Borjas et al., 2008). De igual manera, se observa en otras poblaciones de Sur América estos patrones determinados por una mezcla sesgada por el sexo, pero con variaciones en las proporciones y poblaciones de origen que las conforman, sin embargo, con una tendencia marcada de sesgo entre hombres europeos y

mujeres nativo americanas y africanas, tal y como lo confirman estudios realizados en el análisis del ADNmt y el cromosoma Y en grupos de población Afro-Uruguay (Sans et al., 2002), población local de Colombia (Carvajal-Carmona et al., 2000), y Ecuador (Gonzalez-Andrade et al., 2009; Baeta et al., 2011). También se ha observado la influencia de linajes masculinos europeos en algunas poblaciones consideradas inicialmente como nativas, tal es el caso de los Mapuche, Dieguita y Kolla que habitan en Argentina (Blanco-Verea et al., 2010).

- América Central y el Caribe: En otras regiones del continente americano que fueron colonizados por los europeos como México, el cual en el momento de la conquista presentaba una población nativa de 25 millones de habitantes, sufrió también una dramática reducción del 97% de su población en 1630. En el Caribe, las estimaciones totales para 1492 eran de 1.000.000 habitantes reducidos a 16.000 en el año 1520, lo que explicaría la temprana presencia de esclavos de África en comparación con los demás territorios colonizados. En 1514 la mayor parte de los caciques eran mujeres, ante la patente ausencia de hombres. Otros países de América Central, como Nicaragua y Honduras presentaron cambios demográficos que fueron similares, debido entre otras consideraciones, a la esclavitud de los indígenas centroamericanos y a su exportación a México Perú o Panamá. Entre 1900 a 1930 el crecimiento de estos países latinoamericanos fue del 68%. Dicho incremento vinculado en parte a la inmigración por la demanda de mano de obra, ocurrió sobre todo en los países de la vertiente atlántica (Malamud, 2005).

En resumen, la población total inmigrante del siglo XIX en la mayoría de los países del continente, incidió en un 8-10% sobre la población general de la época, considerablemente más que la población nativa (Morón, 1971), teniendo esto repercusiones en las proporciones de mezcla genética que presentan estas poblaciones en la actualidad. El desplazamiento de los europeos y esclavos africanos hacia el continente americano se observa en la Fig. 2a y 2b respectivamente.

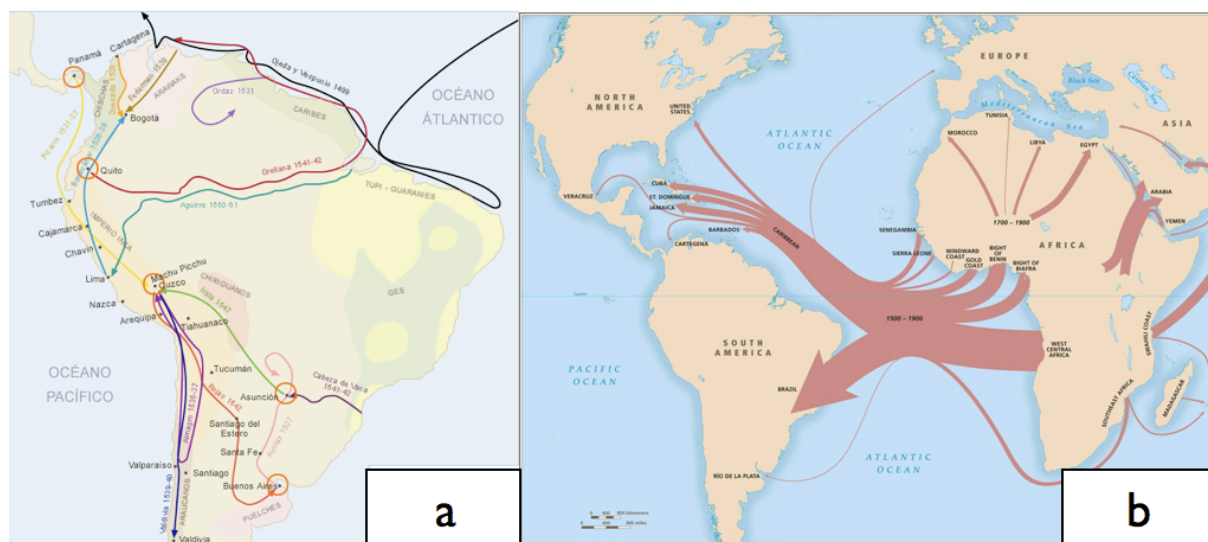


Figura 2. Desplazamientos realizados por europeos (a) y africanos (b) durante el siglo XVI hacia el continente americano. Fuente: *The Trans Atlantic Slave Trade Database* (<http://www.slavevoyages.org/tast/index.faces>).

¿Porqué es importante el estudio de poblaciones multi-étnicas?

El estudio genómico en grupos multi-étnicos, podría contribuir al entendimiento de las variantes genéticas que usualmente llamamos “raras”. Basta con estudiar la historia de la distribución alélica de un gen entre los individuos de diversas poblaciones, para ilustrar cómo las frecuencias de un alelo de riesgo en distintos grupos étnicos, podría tener efectos de gran contraste en los resultados de GWAS que puedan ser obtenidos de un estudio de tipo clínico, por ejemplo. En consecuencia, el poder de predicción de un alelo de riesgo en una población, puede variar considerablemente si se compara luego el efecto de ese mismo alelo en otro grupo, aún y cuando sean empleados el mismo número de casos, controles y niveles de significancia. Por otra parte, el análisis de poblaciones con mezcla podría revelar patrones distintos de LD pudiendo acortar las regiones genéticas de interés, por lo que sería posible incluso con un menor número de muestras detectar una señal de asociación similar, o mayor al de estudios convencionales. La cobertura de los GWAS comercializados en la actualidad, generalmente suele ser baja en poblaciones no europeas, lo que puede ocasionar una inequidad en el empleo de datos para estos análisis. Need y Goldstein en 2009 (Need- Golstein, 2009) realizaron una revisión de gran interés sobre este tema, en la cual recomiendan evitar las posibles disparidades entre poblaciones, sobre todo en lo que a investigación genética clínica se refiere, mediante la inclusión de grupos que presenten una ascendencia diversa, realizando

las correcciones de estratificación correspondientes. De igual forma, también recomiendan en el estudio de genomas completos, incluir todos los grupos ancestrales posibles, con sus respectivos controles de una manera equitativa, aunque esto suponga el empleo de un tamaño de muestra menor.

II.1.b. Grupo Ancestral Biogeográfico (BGA)

En el año 2005, el Instituto Nacional de Investigación del Genoma Humano sugirió emplear el término BGA (Grupo Ancestral Biogeográfico) en lugar de ancestro y raza, entre otras denominaciones ya obsoletas (Halder et al., 2008). BGA, es comúnmente usado para expresar la composición ancestral que es heredada, y su estima a través del ADN ha sido posible debido a las fuerzas evolutivas que han actuado sobre el genoma humano durante aproximadamente 100.000 años, tales como la **mutación, deriva génica, selección natural y flujo genético** (o mezcla), siendo por consiguiente, una medida estrictamente biológica. Las variaciones encontradas en los niveles de BGA entre los individuos de una población, aún teniendo el mismo ancestro genealógico o nivel esperado de grupo ancestral, contribuyen a la sub-estructura de la población y a la valoración del grado de mezcla entre poblaciones. Tanto estas fuerzas evolutivas como la estructura de las poblaciones pueden afectar el equilibrio de Hardy-Weinberg (HWE), en donde se asume que las poblaciones tengan frecuencias alélicas inalteradas, al menos que actúen estas fuerzas evolutivas, siendo esto de particular importancia en el estudio de AIMs (Frudakis, 2008).

El poder estimar el grupo ancestral mediante el estudio de segregación de marcadores autosómicos, permite realizar una medida de la mezcla ancestral de un individuo aportada por todos sus antecesores, diferente ante el estudio de linajes, que sólo representan una estimación de ancestro a nivel de poblaciones. El estudio de BGA en individuos permite reconstruir ciertos aspectos de nuestro pasado evolutivo, así como inferir la dinámica de mezcla y la historia demográfica. De igual forma, el análisis de BGA en individuos es de interés en el estudio de las fuentes de variación genética que puedan contribuir a una cierta enfermedad de riesgo. El poder de estos ensayos clínicos puede llegar a potenciarse significativamente, mediante el empleo de métodos que permitan cuantificar la estructura de la población. Estas bases de

datos de BGA en individuos son de utilidad para la comunidad forense en la inferencia indirecta sobre el aspecto físico de una individuo, así como para el estudio de mapas genéticos en rasgos físicos (Halder et al., 2008).

II.1.c. Determinación de Linajes

El término **linaje** se define como un grupo o *taxa* que comparten un ancestro común, mientras que la **filogenia** se refiere a las estructuras tipo árbol, las cuales representan las relaciones evolutivas entre un conjunto de la *taxa*. La reconstrucción de la filogenia que pueden aportar los marcadores de linaje es posible tanto por vía materna, como paterna - a través del ADN mitocondrial y cromosoma Y respectivamente- dado que constituyen segmentos no recombinantes en el genoma y que poseen una distribución diferencial de las variantes de estos marcadores entre las distintas poblaciones humanas. La filogenia de un *locus* puede proveer información sobre el tiempo y el lugar de su origen. Para poder analizar esto es necesario tener en cuenta las siguientes suposiciones: que la filogenia puede reconstruirse con exactitud y que los movimientos geográficos han sido limitados, por lo tanto que la distribución moderna provee información sobre la distribución antigua (Jobling et al., 2004). Bajo estas y otras suposiciones como la neutralidad, tamaño de población constante y apareamientos al azar, es posible estimar que el tiempo de coalescencia o tiempo del ancestro común más reciente (TMRCA) de un *locus* es proporcional al tamaño efectivo de su población. Asumiendo entonces en humanos, una proporción de sexos de 1:1, se tendría que por cada cuatro copias de un cromosoma autosómico hay tres cromosomas X y hay un cromosoma Y. Por lo tanto, el tamaño efectivo de la población del cromosoma Y se espera que sea un cuarto de la de un cromosoma autosómico, un tercio del correspondiente al cromosoma X y similar al del ADNmt. Se han realizado diversas estimaciones del tiempo de coalescencia de estos marcadores moleculares. La estimación del TMRCA para el cromosoma Y empleando microsatélites es aproximadamente de 46 (16-126) mil años (KY). De manera similar, en estudios posteriores empleando SNPs se estimó un TMRCA de 59 (40-140) KY para este cromosoma. Estos puntos de estimación son recientes comparados con los TMRCA de 177 KY para el ADNmt, 535 y 1860 KY para dos regiones en el cromosoma X y 800 KY para un cromosoma autosómico (Jobling-Tyler-Smith, 2003).

La información obtenida a partir de marcadores de linajes, revela una parte de la historia distinta, pero de complemento a la de marcadores autosómicos. El empleo de AIMs autosómicos para la inferencia del ancestro biogeográfico y el grado de mezcla a nivel de individuos permiten la evaluación de una estructura demográfica “moderna”. Sin embargo, debido a que los *loci* autosómicos residen en las regiones recombinantes de los cromosomas diploides, sus alelos proporcionan una mirada diferente al pasado respecto al ADNmt y el cromosoma Y. Estos marcadores de linajes, a nivel de poblaciones son informativos trazando la filogenia de grupos así como en la inferencia de mezclas (Frudakis, 2008). Un ejemplo que ilustra claramente la información complementaria que existe entre los marcadores de linaje y los autosómicos, es el trabajo presentado por Bedoya y col. en 2005, en el cual se analiza una población americana de ascendencia europea en Antioquia (Colombia), mediante el estudio de ambos tipos de marcadores. Tras el análisis se observó que los marcadores autosómicos confirmaban una ascendencia europea marcada (mayor a la esperada), mientras que por otra parte, se observaba una alta frecuencia de linajes mitocondriales nativos americanos en la población. La interpretación derivada de esta observación, fue que las mujeres nativas estuvieron involucradas en el proceso de mezcla, principalmente durante la fundación de la población, mientras que el flujo genético por parte de hombres europeos ocurrió tanto al comienzo como posteriormente, durante generaciones sucesivas (Bedoya et al., 2009). Por lo tanto, el complemento que ofrecen estos marcadores permite conocer: qué grupos ancestrales participaron en la mezcla, en qué proporción, e incluso una estimación de cuándo ocurrió dicha participación a la mezcla.

Por otra parte, a pesar de la utilidad que ofrecen el ADNmt y cromosoma Y para estudios de grupos ancestrales en poblaciones, su informatividad para inferir rasgos físicos de manera indirecta, resulta muy limitada frente a los AIMs autosómicos (Salas et al., 2006).

- Filogenia del ADN mitocondrial: El ADNmt humano se considera como estrictamente heredado por vía materna. Durante la concepción sólo el núcleo del espermatozoide se une directamente con el óvulo, sin que contribuyan otros componentes celulares del espermatozoide (Butler, 2009). Los polimorfismos encontrados en el ADN mitocondrial son el resultado de

una alta tasa de mutación, posiblemente debido a los diversos eventos selectivos y a las diferencias en la propia estructura de la molécula (por ejemplo, un mecanismo de reparación ineficiente). Sin embargo, actualmente la importancia que se ha dado a la selección positiva sobre el curso de la evolución del ADNmt, es un tema de discusión que requiere de investigaciones que permitan esclarecer si la causa de los patrones evolutivos observados son, en efecto, explicados por esta fuerza (Dowling et al., 2008). La mayor parte del genoma mitocondrial codifica para el producto de 37 genes empleados en el proceso de la fosforilación oxidativa o producción de energía celular. También existe una región llamada “control” que contiene el origen de replicación para una de las hebras del ADNmt, pero no codifica para ningún producto génico. En esta región del ADNmt se encuentran dos de los sitios más polimórficos conocidos como segmentos hipervariables (HVS) I y II. La secuenciación de la región HVSI ha permitido que estén disponibles más de 10.000 secuencias en diversas bases de datos. La ausencia de recombinación en esta molécula y su herencia exclusiva por vía materna permite reconstruir la filogenia para linajes de ADNmt. Sin embargo, la presencia de mutaciones paralelas y reversiones ha afectado este proceso. Existe actualmente más de 500 genomas completos publicados, cuya información en combinación con datos de RFLP y secuencias de la región control, permite la reconstrucción de la filogenia de las principales secuencias de los linajes mitocondriales (Jobling et al., 2004). En el año 2000 se publicó un estudio actualizado sobre la filogenia del ADN mitocondrial, basado en la secuencia completa de 53 individuos de diversos orígenes geográficos, encontrando una gran diversidad de secuencias y posiciones variables fuera del segmento HV (Ingman et al., 2000). A pesar de la alta tasa de mutaciones en el ADN mitocondrial comparado con el ADN nuclear, es posible obtener una filogenia consolidada si se analiza la secuencia completa excluyendo aquellas regiones de alto grado de mutación. Otras de las características de esta molécula que ha permitido la reconstrucción de filogenias es la presencia de múltiples copias de ADNmt por célula en comparación con el ADN nuclear, lo cual ha facilitado el estudio en restos arqueológicos que se encuentran en estado de degradación. Los biólogos evolutivos examinan la variación en la secuencia del ADNmt en relación con

otras especies, en un esfuerzo de establecer posibles parentescos. Un ejemplo de esta aplicación, es en el estudio en Neandertales, donde se determinó que no son antecesores directos de los humanos modernos, esto basado en el análisis de la secuencia de la región control en restos óseos antiguos (Krings et al., 1997).

- Filogenia del cromosoma Y: La información filogeográfica provista por el cromosoma Y tiene su contraparte en el ADNmt, y su comparación ha sido de relevancia particularmente en el estudio del flujo genético sesgado por el sexo que ha acompañado la expansión de los Europeos en América y Oceanía hace 500 años, tal y como se mencionó anteriormente. El interés de estudiar una pieza del genoma que provee información sobre la “mitad” de la población, es debido a que el cromosoma Y es específico del género masculino y de constitución haploide. Éste pasa de padres a hijos, y a diferencia de otros cromosomas escapa en gran parte de recombinación meiótica. Dos segmentos (las regiones pseudo-autosómicas) recombinan con el cromosoma X pero es una región de menos de 3 Mb, sin embargo, la mayor parte de este cromosoma lo constituyen fragmentos no recombinantes. En un cromosoma mayormente no recombinante los haplotipos son la combinación de estados alélicos de marcadores a lo largo del cromosoma, los cuales usualmente pasan intactos de una generación a otra. Estas regiones no recombinantes cambian sólo por mutación y por lo tanto preservan un registro simple en su historia. Mediante el empleo de polimorfismos binarios con bajas tasas de mutación como los SNPs, es posible reconstruir una única filogenia. Al igual que para todas las regiones del ADN - excepto para la región control del ADNmt- la sustitución de bases por mutaciones ocurre a una tasa muy baja como para ser analizadas directamente. Sin embargo, el marco poligénico y la haploidía del cromosoma Y hacen que las mutaciones recurrentes puedan ser identificadas no ambiguamente, y los datos acumulados de los análisis de secuencias pueden proveer información sobre las propiedades de estas mutaciones. Además de los eventos como las mutaciones, la selección es una fuerza potencialmente importante que moldea también la diversidad de haplotipos del cromosoma Y en diversas poblaciones. Debido a la ausencia de recombinación, cualquier evento de

selección podría afectar al cromosoma entero y producir un incremento en la frecuencia de un linaje más rápidamente de lo que puede ser esperado por deriva. El conocimiento de la distribución geográfica de cada linaje es todavía impreciso, algunas poblaciones no han sido aún muestreadas, y los tamaños de aquellas que han sido investigadas, son en pocos casos superiores a cientos de individuos. Por consiguiente, la información ofrecida por los marcadores de linajes deberá ser acompañada por el estudio de *loci* independientes (Jobling-Tyler-Smith, 2003).

II.1.d. Métodos estadísticos para inferir BGA, estructura y mezcla de poblaciones

Para detectar si hay subestructuración en una población se han descrito diversos métodos, de los cuales uno de los más utilizados están basados en el estadístico F_{ST} . Este parámetro deriva de los índices de fijación propuestos por Sewall Wright para medir la desviación de frecuencias de heterocigotos observadas con respecto a las esperadas, de acuerdo con el teorema de Hardy Weinberg (Wright, 1949). El F_{ST} mide la distribución de la variación genética entre sub-poblaciones, es decir, compara la cantidad media de la diversidad genética observada dentro de las sub-poblaciones, con la diversidad genética de la meta-población (grupo de poblaciones conectadas por migraciones). Este estadístico puede ser utilizado como una medida de la proporción de la variación total de las frecuencias alélicas que ocurre entre sub-poblaciones. Si la acción de la deriva genética provoca que las sub-poblaciones sean muy diferentes, esta proporción será grande. Mientras que si una gran cantidad de flujo genético entre sub-poblaciones mantiene su similitud, esta proporción será pequeña. Un inconveniente del estadístico F_{ST} , es que resulta sensible ante cambios en cualquiera de las poblaciones incluidas en el análisis. Tomando esto en cuenta, se desarrolló una extensión de este parámetro que cuantifica el grado de evolución de un locus particular, aislando los cambios en las frecuencias alélicas y permitiendo la especificación no sólo del grado de evolución producida, sino también de la población que ha experimentado ese cambio en particular. Dicha aproximación es conocida como **LSBL** (*Locus Specific Branch Length*) (Shriver MD, 2004). Otros ensayos empleados para estimar una medida de la diversidad suelen estar acompañados de una prueba de significación que demuestre que la probabilidad de subestructuración de la población existente es mayor de la que

podría ser esperada al azar. Un ejemplo de estos métodos son las pruebas de permutación conocidas como **Monte-Carlo**, los cuales analizan aleatoriamente los datos empíricos, y calculan la medida de interés de cada análisis aleatorio, de modo que la medida real de los datos observada es comparada con las medidas simuladas para estimar si es significativamente diferente (Jobling et al., 2004). El grado de estratificación observado en una población, también puede ser cuantificado por el factor de inflación **lambda** (λ) (Devlin-Roeder, 1999). Este parámetro estadístico, mide el grado de “dispersión” entre la media de la distribución del chi-cuadrado obtenido en la población con estratificación y el valor de la media que debería tener en ausencia de ésta. Si el resultado es menor que uno, la distribución es considerada lo suficientemente cercana a lo esperado, y por lo tanto λ es aproximada a uno. La técnica de Control Genómico aplica esta corrección dividiendo el estadístico del chi-cuadrado por su valor λ . Adicionalmente, se han desarrollado diversos programas informáticos específicos para la aplicación de estos y otros métodos estadísticos, por ejemplo, a través del análisis de datos genéticos *multi-locus*, mediante la determinación de frecuencias alélicas ya sean estimadas previamente o durante el proceso de inferencia, para asignar poblaciones de origen a individuos (Rosenberg et al., 2003). Entre los más comúnmente empleados en genética de poblaciones están:

- ADMIXTURE: Este programa permite servir de herramienta en la estimación la máxima verosimilitud de los grupos ancestrales individuales, empleando un algoritmo numérico de optimización rápido en comparación con otros métodos (<http://www.genetics.ucla.edu/software/admixture/>).
- ADMIXMAP: Este programa permite establecer modelos de mezcla de poblaciones, empleando datos genotípicos y demás características de una población que se supone presenta mezcla. En este, los marcadores han sido escogidos con el criterio de tener frecuencias alélicas opuestas entre dos o más poblaciones ancestrales que han contribuido a la mezcla (http://homepages.ed.ac.uk/pmckeigu/admixmap/manual_desc.html).
- EIGENSOFT: Emplea el análisis de componentes principales (PCA) para corregir la estratificación de poblaciones en estudios médicos de asociación

(*EIGENSTRAT*), y para detectar y analizar la estructura de poblaciones (*SMARTPCA*) (<http://genepath.med.harvard.edu/~reich/Software.htm>).

- *PLINK*: Es una herramienta empleada en el análisis de datos de genoma completo, diseñado para realizar estimaciones computacionales en diversas escalas (<http://pngu.mgh.harvard.edu/~purcell/plink/summary.shtml>).
- *STRUCTURE*: En comparación con los otros programas antes mencionados, la flexibilidad de *STRUCTURE* permite combinar diversos análisis, por ejemplo, en la determinación de grupos o *clusters* y la frecuencia de un alelo dado en cada grupo, así como en la determinación de proporciones de cada individuo de corresponder a un determinado grupo ancestral (<http://pritch.bsd.uchicago.edu/software.html>).

II.1.e. Análisis de genomas completos en poblaciones humanas

Actualmente, son cada vez más las tecnologías de alto rendimiento que incrementan considerablemente la cobertura genómica y el número de poblaciones para proyectos como HapMap y HGDP. Los análisis de genomas completos han sido utilizados en diversos estudios de variabilidad genética, en los que se pretende inferir patrones de variación de todo el genoma y estudiar la subestructuración de las poblaciones, así como para evaluar el ancestro biogeográfico, determinar el grado de mezcla entre poblaciones, e identificar factores genéticos de riesgo a enfermedades complejas (Sanchez-Diz-Ramos-Luis, 2010). Existen diversos estudios en donde a través del análisis de genomas completos, ha sido posible tener un mejor conocimiento sobre las fuerzas selectivas que han actuado a lo largo del tiempo en diversas poblaciones del mundo. Por ejemplo, en el estudio de Li en 2008 (Li et al., 2008), en el cual tras el análisis de 7 poblaciones definidas como África, Medio Este, Europa, Asia del Este y del Centro-Sur, Oceanía y América, se observó una evidente variación genética entre las poblaciones, indicando la presencia de subestructuración consecuencia de la deriva aleatoria de *loci* neutros, aunque considerando también que algunas regiones del genoma pueden haber experimentado una divergencia acelerada, debida a selección local. Esta observación ha sido posible realizarla, incluso a mayor escala, entre diversas regiones geográficas dentro de la población

Europea (Novembre et al., 2008). Posteriormente se encontraron evidencias que corroboraron la presencia de un flujo genético desde África, considerando una hipótesis de refugio glacial debido a la presencia de un gradiente de diversidad haplotípica norte-sur (Auton et al., 2009). Todas estas observaciones han sido posibles gracias al estudio de genomas completos. Sin embargo, hay que mencionar que, entre las principales limitaciones que pueden tener el empleo de *arrays* a gran escala como el 5.0 de *Affymetrix* empleado en este último estudio, está el hecho de que pueden no poseer la suficiente potencia para detectar haplotipos en ciertos grupos como en la población del continente americano.

II.2. External Visible Characteristics (EVC)

Como se ha mencionado anteriormente, las EVC de interés forense están referidas a todas aquellas características humanas determinadas por *loci* funcionales, basados en la expresión de un rasgo físico que pueda aportar información sobre la apariencia de una persona, siendo de utilidad para la investigación criminal, así como para el estudio de restos antiguos y desaparecidos en catástrofes naturales. Desde el punto de vista genético, las características físicas externas son generalmente complejas. Los rasgos físicos complejos están definidos como aquellas características que se encuentran bajo la influencia de varios genes así como debido a factores ambientales. Aunque algunas EVCs fueron inicialmente considerados como rasgos simples causados por únicos genes siguiendo una herencia Mendeliana, actualmente se conoce que en la mayoría de los casos, más de un factor genético debe estar involucrado, así como diversos factores ambientales.

II.2.a. Quantitative Trait Locus (QTL)

En la literatura actual sobre EVCs, es muy común relacionar este término con características “fenotípicas”. Sin embargo, es importante tomar en consideración que el empleo del término fenotipo bajo este contexto de predicción de EVCs, debe ser manejado con especial precaución, ya que por definición el fenotipo es la expresión del genotipo en un determinado ambiente, y se refiere generalmente a rasgos cualitativos. Las características físicas que son tratadas en este estudio como la pigmentación, están mejor definidas como de carácter cuantitativo o QT (*Quantitative*

Trait), que se refieren a rasgos que pueden variar en grados continuos y que pueden ser atribuidos a efectos poli-génicos (producto de dos o más genes) y a su interacción con el medio ambiente. Sin embargo, en algunos estudios (incluso en pigmentación) por convención se suelen asignar clases fenotípicas a características que son en realidad de tipo cuantitativas, para así poder dirigir el análisis estadístico correspondiente (Sturm-Frudakis, 2004; Sturm-Larsson, 2009). Por su parte, los QTLs (*Quantitative Trait Locus*) son secuencias de ADN que contienen o que están vinculados a genes relacionados con un carácter cuantitativo (Strachen-Read, 1996). La determinación de QTLs se realiza mediante la correlación entre la segregación de rasgos fenotípicos en la descendencia de cruces con la segregación de marcadores genéticos polimórficos. El estudio realizado por la genética cuantitativa, descompone la varianza total fenotípica en un componente genético y un componente ambiental. La proporción de la varianza en un rasgo que puede ser explicado por factores genéticos es conocido como grado de herencia. El componente genético puede a su vez descomponerse en aditivo, dominante y epistático (Jobling et al., 2004). La identificación de los genes que contribuyen a los rasgos físicos complejos representan un desafío para la comunidad científica (Glazier et al., 2002; Pulker et al., 2007) debido a factores como:

- Interacciones de tipo gen-gen: El significado de las interacciones entre genes para la producción de un efecto sobre el fenotipo se conoce desde que en 1909 Bateson (Bateson, 1909) describió un efecto de tipo enmascarado de un alelo en un gen sobre otro. Posteriormente, la epistasis fue definida como una desviación en los efectos aditivos de las variantes en distintos *loci* con respecto a su contribución a un rasgo fenotípico determinado. En el presente, la epistasis como interacción genética, es entendida como una desviación del modelo lineal que describe cómo varios factores predicen un efecto fenotípico final (Pospiech et al., 2011). Este fenómeno suele estar representado a través de tablas de penetrancia (cuyos valores son generalmente bajos para estos rasgos) donde se muestran “modelos de heterogeneidad”, que corresponden a la frecuencia con la que los individuos que poseen un genotipo de predisposición particular, manifiestan un fenotipo dado (Cordell, 2002). Aunque en la actualidad existe una variedad de métodos estadísticos disponibles para evaluar la interacción gen-gen, aún

es difícil indicar qué herramienta es la más efectiva dada las limitaciones existentes en cuanto al poder estadístico. Las regresiones logísticas o lineales, han sido durante largo tiempo las pruebas más comúnmente utilizadas para la evaluación de efectos de tipo epistático. Sin embargo, es bien sabido que la dispersión de datos y la multi dimensionalidad puede provocar el incremento del error tipo 1 (la obtención de falsos positivos) cuando se aplican métodos de regresión. En el año 2001, se dio a conocer el método MDR (*Multifactor Dimensionality Reduction*), el cual es de tipo no paramétrico, y no asume ningún modelo de herencia genética, lo que hace de éste una herramienta efectiva para la determinación de interacciones entre genes. Este método permite también la reducción de datos genotípicos multi *locus* en una sola dimensión, así como una evaluación del modelo mediante el empleo de validaciones cruzadas y de pruebas de permutación (Branicki et al., 2009).

- Pleiotropía: Es el fenómeno mediante el cual un mismo gen da origen a múltiples efectos en el fenotipo (Tamarin, 2001).
- Heterogeneidad de *locus*: Se define como la presencia de características aparentemente similares por las cuales, las evidencias genéticas indican que diferentes genes o diferentes mecanismos genéticos se encuentran involucrados en diversos pedigríes de un mismo fenotipo. En otra definición, la heterogeneidad genética se refiere a la presencia de diversas alteraciones genéticas las cuales originan un mismo fenotipo (Rigomar Rieger et al., 1992).
- Fenocopia: Es el resultado de una deficiencia en la dieta o de un trauma producido por efectos del medio ambiente, y que careciendo de un genotipo dado es posible observar la expresión de un carácter independientemente de su variable genética (Tamarin, 2001).
- Aspectos evolutivos de la variabilidad de rasgos físicos y su interacción con el medio ambiente: La variabilidad de las características físicas entre las diversas poblaciones humanas, ha estado en parte bajo la influencia de diversos aspectos de nuestro pasado evolutivo tales como la diferenciación y dispersión temprana de las poblaciones humanas; la colonización inicial de

nuevos ambientes, la adaptación a los cambios climáticos, el desarrollo de la agricultura, las expansiones demográficas recientes, los cambios en la dieta, el contacto cercano con otras especies animales, la mezcla reciente entre poblaciones con historias diferentes, así como las migraciones recientes. Las características fenotípicas se han establecido en poblaciones humanas durante largos períodos, cuando el tamaño de las poblaciones era aún pequeño, la migración entre los diferentes grupos era mucho menor que en la actualidad y nuestra habilidad de influir sobre el medio ambiente por el empleo de tecnologías era más limitada. Las intervenciones de tecnologías modernas parecen indicar por una parte, que la selección natural podría ser un factor que influirá en la frecuencia de muchos rasgos físicos cada vez en menor medida, y que la especificidad de los grupos geográficos quedará cada vez más reducida por la migración y la mezcla de poblaciones. Cualquier rasgo cuantitativo que esté codificado genéticamente y bien diferenciado entre grupos, podría sufrir alteraciones en aquellas poblaciones geográficas donde ocurra mezcla (Jobling et al., 2004).

II.2.b. Tecnologías para la identificación de genes asociados a rasgos complejos

Para la identificación de genes relacionados a rasgos complejos, se ha sugerido el empleo de los siguientes criterios de análisis: (a) asociación, (b) ligamiento, (c) secuenciación y (d) estudios funcionales (Glazier et al., 2002).

Por su parte, los avances recientes en las tecnologías basadas en el genotipado de *microarrays*, han permitido evaluar en paralelo más de un millón de marcadores genéticos (usualmente SNPs) junto con los avances en estudios de asociación, proporcionando herramientas de gran valor en la identificación de genes involucrados en rasgos complejos, incluyendo algunos EVCs (Kayser-Schneider, 2009). El diseño de un estudio de asociación para rasgos complejos, requiere generalmente de: un gran número de muestras, una definición establecida del fenotipo así como de las poblaciones de estudio, y la decisión de emplear genes candidatos o análisis de todo el genoma (**WGS Whole Genome Scans**). Los WGS tienen la ventaja de estar libre del sesgo por genes específicos, pero la desventaja de ser una tecnología costosa. Por el contrario, si se escoge el análisis de **genes**

candidatos, estos deben ser seleccionados de forma apropiada a través del análisis de rutas metabólicas, genómica comparativa, perfiles de expresión genética, o explorando AIMs (debido a que la marcas de la selección pueden proporcionar pistas sobre los genes involucrados en rasgos complejos) (Pulker et al., 2007). El enfoque de genes candidatos contribuye a delinear las rutas bioquímicas involucradas en la expresión del fenotipo, pero no permite observar los efectos significativos en muchos genes en poblaciones humanas. Una de las razones de esta limitación puede ser debido a que muchos genes candidatos han sido identificados como el resultado de fenotipos observados en estudios funcionales con genes *knockouts* en ratones, y que las distancias evolutivas entre estas especies pueden ser suficientemente grandes para que ocurriese una divergencia funcional. Y otra posible razón, es que en humanos algunos genes que han sido escogidos como candidatos a contribuir a una determinada variación “común” en un rasgo, se deben a mutaciones en esos mismos genes que pueden producir un fenotipo “alterado”, y estas observaciones no siempre pueden ser extrapoladas de forma directa (Jobling et al., 2004).

Aunque los estudios sobre los genes relacionados con la predicción de EVCs han sido previos al el empleo de los GWAS en la actualidad, es esta última tecnología quien ha permitido conocer nuevos genes y marcadores con una asociación significativa a las diversas EVCs. Una vez que se ha identificado el intervalo donde se encuentra contenida la variación asociada al fenotipo, es posible refinar luego el análisis obteniendo un intervalo más corto con menos genes empleando análisis de ligamiento, pero usualmente todavía estos intervalos contienen muchos genes. Debido que algunos marcadores genéticos incluidos en estos *arrays* representan polimorfismos localizados mayormente en las regiones no codificantes del genoma, hay que tomar en cuenta también, que resulta poco probable que los estudios de GWAS puedan identificar las variables causales, las cuales deben ser estudiadas en investigaciones posteriores. No obstante, si la asociación de un marcador no causal a un rasgo físico es suficientemente fuerte, y éste marcador puede explicar una larga proporción de la variación del rasgo, entonces esta asociación es de gran valor para el empleo de pruebas de predicción de EVCs. Sin embargo, si una determinada EVC está restringida a individuos de una región geográfica específica y los marcadores en el ADN predictivos están asociados pero no son causales, estos ensayos de predicción de características físicas externas deberían combinarse con una prueba basada en el

empleo de marcadores de ADN que permitan la inferencia de grupos ancestrales, para poder así evitar interpretaciones engañosas. Por el contrario, no se precisa de un ensayo de predicción de grupos ancestrales si los marcadores predictivos de rasgos físicos son causales y funcionalmente responsables de la EVC de interés (Kayser-Schneider, 2009).

II.2.c. Algunos EVCs estudiados con aplicación forense

- Sexo: Durante la década de los 90's, se implementó por primera vez en un test genético con fines forenses la determinación de variantes en el gen de la amelogenina para la inferencia del sexo (Akane et al., 1991), lo cual no sólo permitió obtener información sobre la apariencia física *a grosso modo*, sino que en la actualidad forma parte de los diversos *kits* comerciales de identificación individual con STRs (Sullivan et al., 1993); (Mannucci et al., 1994). Sin embargo, este ensayo puede estar sujeto a ciertos errores en la determinación del sexo, ya que se ha descrito la presencia de deleciones en el cromosoma Y que pueden incluir el *locus* de la amelogenina, pudiendo esto ocurrir entre los diversos grupos de poblaciones geográficas, por lo que se ha sugerido adicionar nuevos marcadores genéticos, de tal forma que sea posible confirmar dicha predicción (Kayser-Schneider, 2009).
- Estatura: La altura en edad adulta, representa un rasgo complejo pero que puede ser clasificado a través de clases discretas (empleando por ejemplo proporciones). Este rasgo posee un factor de herencia de un 80% según estudios de ligamiento realizado en diversas familias (Visscher et al., 2008), en los cuales se favorece la presencia de factores genéticos involucrados en la altura. Este rasgo es en gran parte el resultado del efecto aditivo de un gran número de genes, así como también se encuentra bajo la influencia de otros genes que aún no han sido detectados. Algunos determinantes de la estatura baja se encuentran asociadas a mutaciones en los genes relacionados con la expresión de la hormona del crecimiento (OMIM 300582, 31286, 604251) (Pulker et al., 2007). En el año 2008, tres estudios de GWAS realizados en busca de factores genéticos relacionados con la altura humana, revelaron la presencia de 54 variables genéticas validadas con una asociación

estadísticamente significativa a este rasgo (Gudbjartsson et al., 2008; Lettre et al., 2008; Weedon et al., 2008), incluyendo algunas variables detectadas en estudios anteriores (Weedon et al., 2007; Sanna et al., 2008). Sin embargo, éstas sólo explican una pequeña proporción en la variación de la altura (0,4-0,8 cm). Por otra parte, se espera que los demás factores genéticos a ser investigados tengan efectos en el fenotipo incluso menor. Por lo tanto, el conocimiento actual sobre los componentes genéticos que afectan este rasgo son explicados por múltiples variantes genéticas, con pequeños efectos individuales cuya utilidad es aún limitada para la investigación forense (Kayser-Schneider, 2009).

- Morfología facial: La posibilidad de medir objetivamente la morfología facial para definir grupos fenotípicos permite obtener una primera vista de la variación de este rasgo dependiendo de su definición como de tipo continuo o discreto. La similitud observada entre individuos monocigóticos y dicigóticos indica que existe una alta herencia para estas características. Algunos estudios han explorado el efecto genético a nivel cuantitativo respecto a la variación en las dimensiones de los rasgos antropométricos en diversas poblaciones. Se ha encontrado un rango en el grado de herencia entre 0,48-0,90 para la forma de los ojos, y de 0,5-0,74 para la forma de la nariz, lo que sugiere una importante contribución genética a las diferencias encontradas entre las dimensiones de estos rasgos. Igualmente en dichos estudios, se ha sugerido como genes candidatos determinantes de la morfología ocular a *TCOF1* y *MSX2* entre otros (Im et al., 2010). Por otra parte, se han descrito otras características aunque no directamente relacionadas con la morfología facial pero de interés en la descripción de un rostro, y que además obedecen a una herencia de tipo Mendeliana como es el caso del hoyuelo del mentón (OMIM 126100), la presencia de pelo en las orejas (OMIM 139500), la forma del lóbulo de la oreja (OMIM 128900), y el “pico de viuda” (OMIM 194000) (Pulker et al., 2007).
- Edad: Otra característica que podría ser de gran valor en la predicción de la apariencia física a partir del ADN en la investigación forense es la edad. Se ha sugerido el empleo de dos enfoques para la predicción de la edad, uno de

ellos basado en la acumulación de deleciones en el ADNmt, así como la reducción de las regiones teloméricas que ocurren a lo largo del tiempo en el genoma de un individuo. Sin embargo, el valor práctico de esta inferencia se ha visto limitada por factores como los coeficientes de correlación hallados (que no superan el 0,90), y porque estos métodos determinan la edad biológica y no cronológica (Meissner-Ritz-Timme, 2010). Estudios de asociación en genomas completos han evaluado los cambios en la expresión genética respecto a la edad, así como los patrones observados en la metilación del ADN, los cuales han proporcionado información de interés en la búsqueda de nuevos biomarcadores que permitan inferir esta característica tan compleja (Lu et al., 2004; Teschendorff et al., 2010). Posiblemente, uno de los estudios que más se aproxima a una inferencia más exacta de la edad, se ha realizado en tejido sanguíneo, basado en el conocimiento previo sobre el decrecimiento de las células T (y por lo tanto de los *TRECS T-cell receptor excision circles*), conforme aumenta la edad. Este estudio demostró que la cuantificación de las moléculas de ADN en los *TRECS* puede ser empleado para estimar la edad a partir de muestras sanguíneas con un error estándar de ± 9 años, así como también se ha evaluado la predicción de la edad en grupos categóricos de 20 años con valores de área bajo la curva AUC entre 0,89-0,97, evaluando también la sensibilidad del ensayo en muestras afectadas por la degradación (Zubakov et al., 2010).

- Morfología del cabello: La morfología del cabello es una de las características, junto con la pigmentación más visibles en humanos, y es particularmente diversa entre individuos de población Europea. El grado de curvatura del cabello posee una cierta correlación entre la distribución de la queratina del cabello y el tipo celular de la fibra del cabello. Estudios recientes han identificado en población asiática alelos específicos de estas regiones geográficas en los genes *EDAR* y *FGFR2* como asociados con el grosor y la forma lisa del cabello, sugiriendo además que esas variantes surgieron después de la divergencia entre europeos y asiáticos (Fujimoto et al., 2008; Mou et al., 2008). Posteriormente, en población con ascendencia europea, se han identificado a través de GWAS algunas variantes en el gen *TCHH* (*Trichohyalin*) el cual se expresa en el folículo capilar, explicando un 6%

de la varianza de la morfología de este rasgo (Medland et al., 2009). Por su parte, existe información limitada respecto a la orientación en la que crece el cabello, como algunos estudios funcionales que proponen el gen *DAPT* como asociado a cambios en la orientación de los estereocilios (microvellosidades) en células del folículo capilar de ratones (Zhao et al., 2011).

Además de los EVC antes mencionados, existen otras características que aunque no forman parte de la apariencia visual de un individuo, podrían ser también de interés en la investigación criminal, como la habilidad de las manos (diestro, zurdo o ambidiestro) o el grado de miopía, sin embargo, su investigación genética ha sido muy limitada, por lo que en la actualidad sólo podrían ser considerados como características potencialmente de interés forense. (Pulker et al., 2007). Por otra parte, además de las características ya establecidas en la rutina forense, otra de las EVCs que actualmente se está comenzando a aplicar en algunos casos de investigación criminal, es la inferencia de la **pigmentación humana**, tema a desarrollar en el presente estudio.

II.2.d. La pigmentación humana

La pigmentación de piel, cabello y ojos, representan uno de las EVCs que mejor se han identificados en humanos, sin embargo, se trata de un rasgo poligénico complejo (Badano-Katsanis, 2002; Sturm-Frudakis, 2004). Existe un alto grado de variación en el color de piel, cabello y ojos, entre individuos tanto de una misma población, como entre diversas poblaciones. (Frost, 2006; Frost, 2007; Sturm, 2009). El pigmento más importante que influye en la coloración de los diversos rasgos visibles humanos es la melanina (Sturm et al., 1998). Esta sustancia granular es producida por células especializadas, llamadas melanocitos, y se encuentra concentrada en vesículas llamadas melanosomas. El número de melanocitos es aproximadamente el mismo entre diferentes individuos y las variaciones en las coloraciones resultan de las diferencias en el número, tamaño y distribución de los melanosomas (Sturm-Frudakis, 2004). Por otra parte, el tipo de melanina dentro de los melanosomas también tiene una gran influencia sobre la pigmentación. En cabello y piel, los melanosomas son transferidos desde los melanocitos a los queratinocitos

circundantes. En contraste, en los ojos los melanosomas están retenidos en los melanocitos del iris (Sturm-Larsson, 2009).

Los genes implicados en la pigmentación humana participan en diversas rutas bioquímicas, incluyendo (a) aquellas para la formación del complejo de la enzima tirosinasa en la superficie interna del melanosoma, (b) las de regulación hormonal y señalización (*ASIP*, *MC1R*, *POMC*, *OA1*, *MITF*), (c) las de diferenciación y migración de los melanocitos (d) rutas de proteínas dentro del melanosoma (*TYRP*, *TYRP1*, *DCT*, *SILV*, *OCA2*, *MATP*), y (e) aquellas que corresponden al propio transporte del melanosoma a los queratinocitos (*MYO5A*, *RAB27A*, *HPS1*, *HPS6*) (Frudakis et al., 2003b; Sturm-Frudakis, 2004).

II.2.d.1. Patrones de distribución de la pigmentación humana a nivel mundial

En el caso de la **pigmentación en piel**, se ha observado que existe una distribución geográfica de este rasgo de tipo no azarosa, particularmente en la población nativa a lo largo del globo terráqueo (Fig.3a). De acuerdo con esta distribución, se ha descrito un patrón con la coloración de piel más oscura en los trópicos y una pigmentación más clara en latitudes de zonas norte. Se ha propuesto un gran número de posibles explicaciones sobre esta variación. Sin embargo, al igual que para otras características fenotípicas adaptativas, existen poblaciones que no se ajustan a este patrón. De hecho, existen estudios que mediante el empleo de programas como *STRUCTURE* ha sido posible demostrar que la relación entre las poblaciones ancestrales y los tipos de pigmentación en piel no sólo se deben a los patrones de distribución geográfica, sino que poseen una base en aquellos genes que son causales de esta característica (Lao et al., 2007), y que además, las tonalidades de piel clara no son sólo el resultado de una pérdida presión ambiental, sino que posee también un valor adaptativo (Izagirre et al., 2006). Si se considera la pigmentación de piel como un rasgo cuantitativo, la proporción de variación entre individuos es alta, si ésta se compara con los resultados obtenidos con marcadores neutrales.

Éstas diferencias son posiblemente el resultado de la adaptación a otras fuerzas selectivas, y por lo tanto un escaso reflejo de las relaciones entre las poblaciones (Jobling et al., 2004). Muchos antropólogos han intentado encontrar una

respuesta sobre el probable estado ancestral del color de piel y su posterior adaptación ante la exposición a radiación UV (UVR), especialmente ante UV-B (cuya longitud de onda oscila entre 290-315 nm), la cual representa la forma de radiación ultravioleta más energética que normalmente alcanza la superficie de la tierra. Para poder proteger la piel del daño producido por la UVR, el proceso de producción de melanina en los humanos modernos se ha favorecido en gran medida.

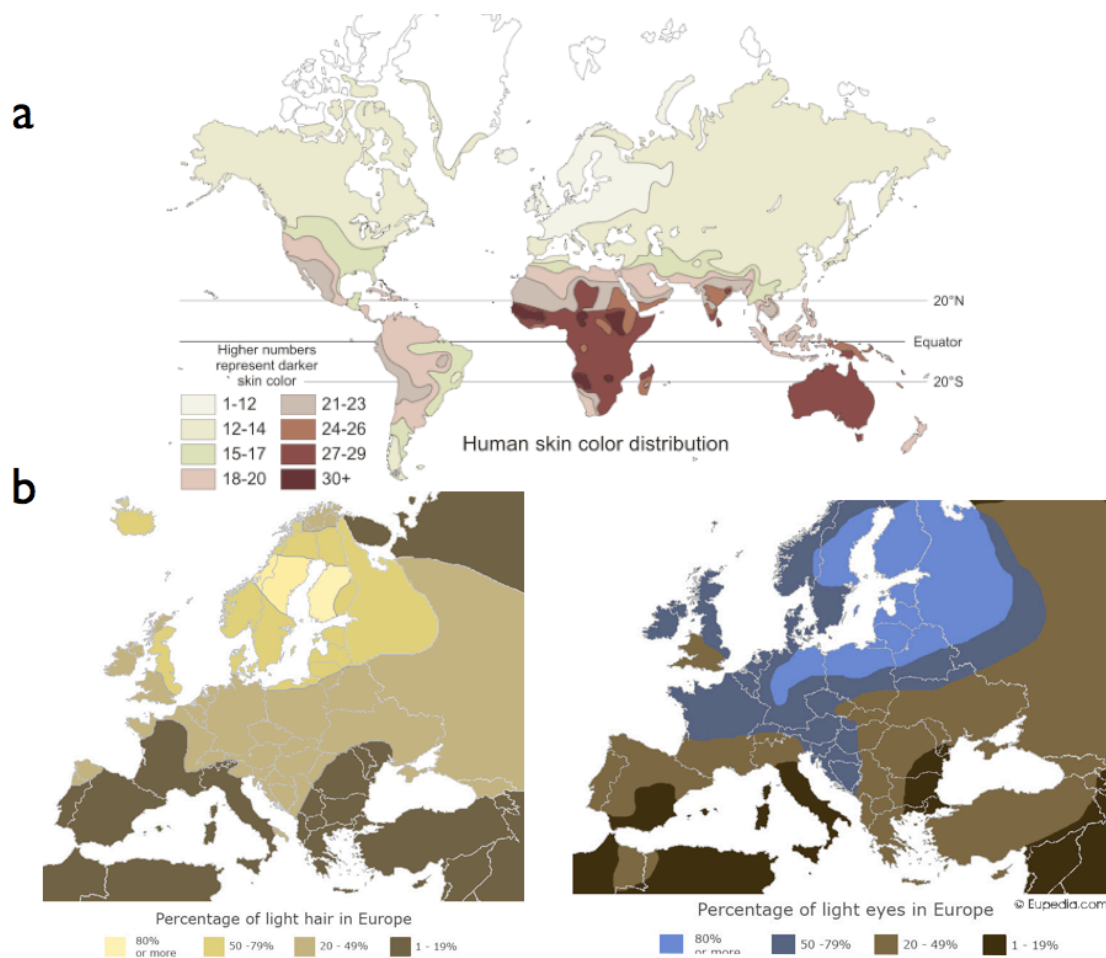


Figura 3. Patrones de distribución geográfica de piel basado en la escala cromática de Von Luschan(a) Fuente: Datos de Biasutti, 1959 (Biasutti, 1959). Patrón de distribución en Europa de pigmentación en cabellos y ojos (b). Fuente: Parra, 2004 (Parra et al., 2004)

La melanina posee un doble papel como protector de los UVR. En primer lugar, reduce la cantidad de radiación que entra en las capas internas de la epidermis mediante su absorción o dispersión. Y en segundo lugar, en virtud de su naturaleza química, actúa como un filtro absorbiendo los productos derivados de los daños causados por la UVR, que puedan ser tóxicos o cancerígenos. Adicionalmente, existen algunas evidencias paleontológicas que indican que la necesidad de producción de vitamina D (puesto que la UVR resulta esencial para la síntesis de

vitamina D, necesaria para la regulación de la absorción del calcio), la fotoprotección, las adaptaciones culturales, y posiblemente la selección sexual han desempeñado un papel importante en la distribución de diversos patrones de colores de piel (Jablonski-Chaplin, 2000; Jobling et al., 2004). Se espera que si la pigmentación obedece a un rasgo adaptativo, es posible determinar la presencia de una selección positiva en algunos de los genes involucrados, tal es el caso de *TP53BP1* en población africana, y *TYRP*, *TYR*, *MATP* (*SLC24A5*) en población caucásica (Izagirre et al., 2006), mientras que en otros estudios se ha observado que tanto *OCA2* como *ASIP* pueden jugar un papel compartido contribuyendo a la presencia de pieles claras y oscuras a nivel mundial (Norton et al., 2007).

La variabilidad de **pigmentación en color de ojos y cabello** presenta un patrón de distribución algo diferente al encontrado en piel. Dicha diferencia radica en que se ha descrito una particular variabilidad de estos rasgos sólo en población nativa Europea, especialmente en zonas del norte y del este (Fig. 3b), sin embargo, se desconoce la razón de esta peculiaridad respecto a otras regiones geográficas en donde no se observa tal variación. La cantidad de genes involucrados y su origen independiente desarrollado en un corto periodo de tiempo en la evolución humana, han indicado que estos rasgos pueden ser el producto de algún tipo de selección, como la sexual. Dicha selección ha podido actuar sobre estos rasgos debido a que se conoce para otras especies que este tipo de selección favorece en gran medida aquellas coloraciones que puedan surgir espontáneamente, y ser una novedad ante el rasgo común (Frost, 2006). En otros estudios en población caucásica, se ha indicado que la exposición a niveles altos de estrógeno prenatal es más frecuente en individuos de cabellos claros y ojos no marrones respecto al resto (Manning et al., 2004). Se cree que la mutación responsable para la coloración de ojos azules probablemente se originó en las regiones cercanas al noreste de la región del Mar Negro, en donde ocurrió una gran migración de la agricultura hacia el norte de Europa la cual tuvo lugar en el período Neolítico, hace 6-10000 años (Cavalli-Sforza et al., 1994).

Aparte de la coloración de cabello pelirrojo, para el resto de las coloraciones del cabello, así como para la de ojos, no existen evidencias contundentes que indiquen que existe un fuerte ligamiento genético de estos rasgos al color de piel.

Esto parece indicar que las fuerzas selectivas han actuado simultáneamente sobre estos rasgos y para estas poblaciones en concreto, mientras que ha estado ausente en latitudes similares del norte de Asia y América (Frost, 1994).

II.2.d.2. Melanogénesis

La melanina es un biopolímero inerte que absorbe luz, de tamaño variable y resulta muy resistente a la degradación. La estructura química exacta de las diferentes clases de polímeros de melanina es compleja, sin embargo, sólo se conocen dos de los principales tipos de melanina: la eumelanina y la feomelanina, los cuales son responsables de las coloraciones oscuras (negro y marrón) y claras (amarillo y rojo) respectivamente. Las rutas bioquímicas que corresponden a la síntesis de los diversos pigmentos de melanina se conocen como melanogénesis. Estas rutas se basan en las distintas reacciones químicas entre la tirosina, dopa y cisteína que tienen lugar en el melanosoma y que dan como resultado la formación de los pigmentos, a través de una ruta biosintética bifurcada (Ito, 2003). Cuando la tirosina es oxidada por la enzima tirosinasa (TYR), se produce la dopaquinona (DQ) como molécula intermediaria. En ausencia de la cisteína, la DQ experimenta una adición intramolecular (acumulación) produciendo ciclodopa, con un intercambio redox entre la ciclodopa y DQ dando origen a las moléculas de dopa y dopacroma. La dopa es un sustrato que estimula a la TYR para incrementar la producción de más DQ, y por lo tanto la tasa de melanogénesis. Por su parte, la dopacroma se descompone para dar origen mayormente a 5,6-dihidroxyndol (DHI), y con la acción catalítica de la dopacromo tautomerasa (DCT) también produce 5,6-dihidroxyndol-2 ácido carboxílico (DHICA). Estos compuestos son posteriormente oxidados por las enzimas TYR y la TYRP-1 (proteína relacionada con la tirosinasa), para finalmente producir eumelanina (Fig. 4- izquierda).

En otra ruta, en presencia de cisteína la DQ se conjuga con ésta originando a la 5-S-cisteinildopa, y en menor medida a la 2-S-cisteinildopa. Estas moléculas son posteriormente oxidadas para dar origen a intermediarios de benzotiacinas, que son incorporados al polímero rojo-amarillo de la feomelanina (Fig. 4- derecha). Se conoce poco sobre el proceso de regulación catalítica involucrado en la generación de feomelanina, pero se estima que la adición de cisteína a DQ es un proceso rápido, y

que éste continúa siempre que haya disponible cisteína en el melanosoma. La oxidación de las cisteinildopas y la incorporación a la feomelanina continúa siempre que esté disponible la cisteinildopa. El agotamiento de la cisteína y la cisteinildopa en el melanosoma promueven el comienzo de la ruta eumelanogénica, con el depósito de eumelanina sobre la feomelanina producida. Por lo tanto, cada melanocito puede producir ambos tipos de pigmentos, y cuando esto ocurre este fenómeno se denomina melanogénesis mixta. Los ratios entre las dos formas de melanina pueden variar ampliamente entre individuos, lo cual se refleja entre las diversas tonalidades presentes en coloración de ojos, cabello y piel (Sturm et al., 1998).

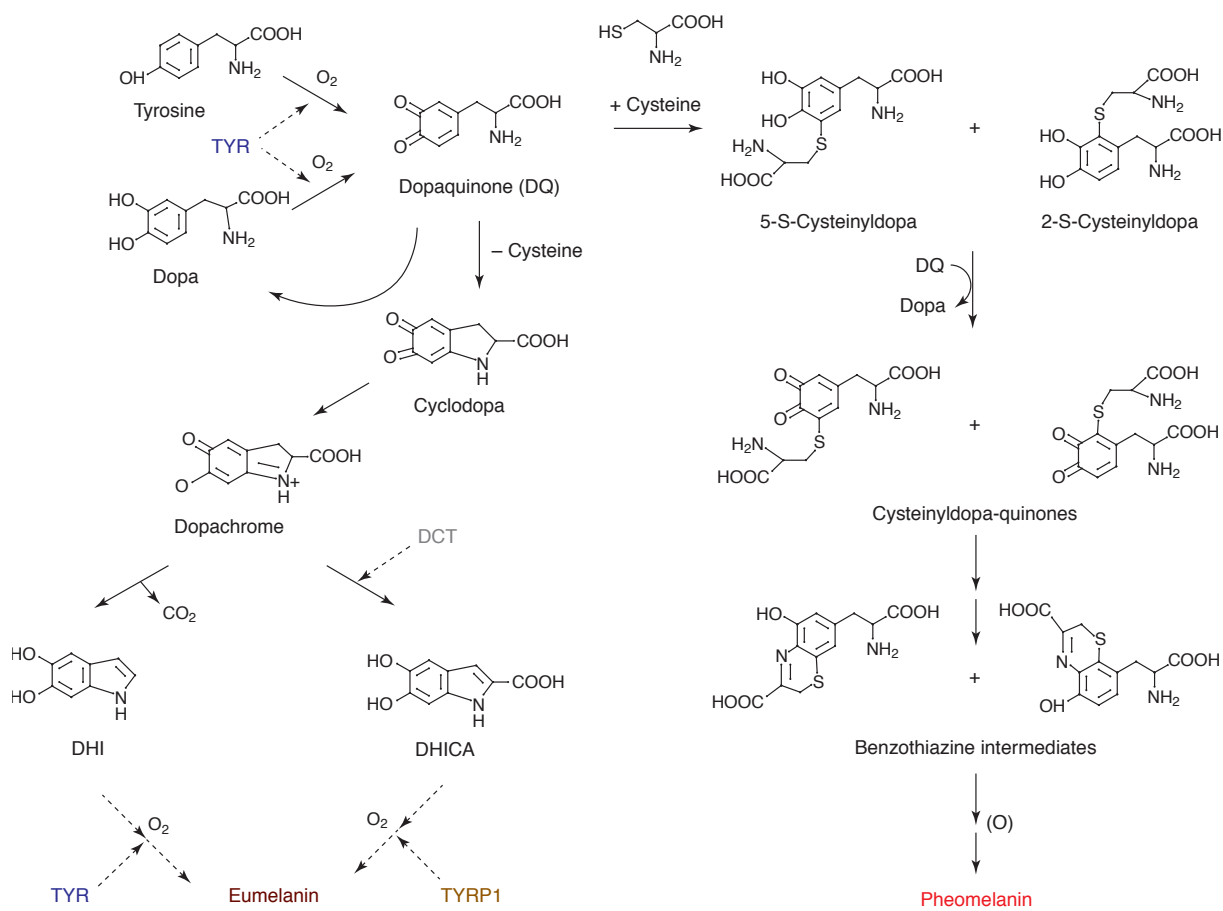


Figura 4. Formación de dos de los principales pigmentos de la melanina: Eumelanina (izquierda), y Feomelanina (derecha). (Sturm, 2004)

Adicionalmente, existen una serie de proteínas específicas de los melanosomas también responsables de este proceso (además de la TYRP1, 2, y la DCT) las proteínas de membrana transportadoras P (OCA2), MATP/SLC45A2 y SLC24A5, implicadas en el control del pH en el melanosoma, la osmolaridad y el contenido de calcio respectivamente (Puri et al., 2000; Ancans et al., 2001; Newton et al., 2001; Lamason et al., 2005). Por su parte, el proceso de formación y transporte puede a su vez estar

estimulado por diversas rutas de señalización que llegan desde la membrana plasmática del melanocito, siendo una de las más relevantes la mediada por el receptor-1 de la melanocortina (MC1R), el cual permite la reducción de especies de oxígeno reactivas y estimula los mecanismos de reparación del ADN (Fig. 5). Dicho receptor es estimulado por α -MSH (hormona estimulante del melanocito) y ACTH (hormona adenocorticotrófica) e inhibido por ASIP (proteína de señalización *agouti*). Entre otra de las rutas, también está la mediada por el factor de transcripción de la microftalmia (MITF), el cual al unirse a la región promotora del ADN, estimula la transcripción de genes que codifican proteínas implicadas en la ruta de la eumelanogénesis, así como también promueve el incremento en el transporte, número y tamaño de los melanosomas (Lin-Fisher, 2007; Schiaffino, 2010). Por su parte, la ruta mediada por el receptor del melanosoma acoplado a la proteína G (GPCR) OA1 (proteína del albinismo ocular tipo 1), actúa como sensor de la maduración del melanosoma, y es activada por moléculas, posiblemente de la ruta de biosíntesis de la melanina como L-DOPA.

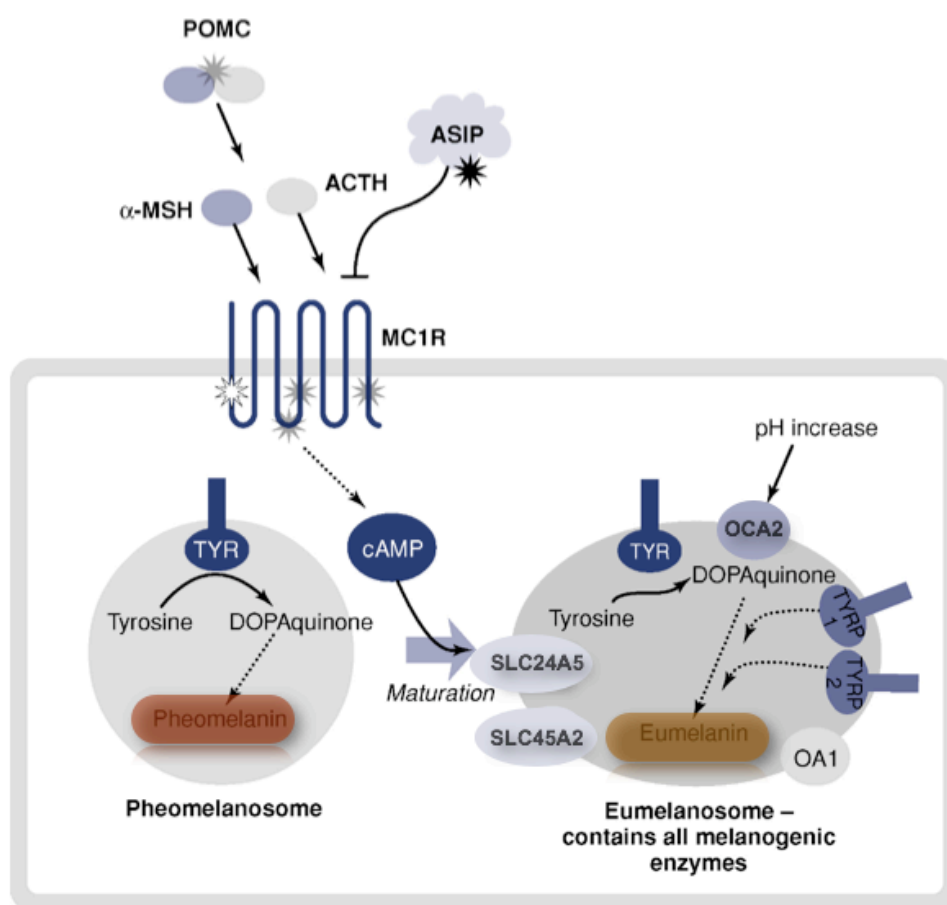


Figura 5. Regulación de la melanogénesis por estimulación del receptor de la melanocortina-1 (MC1R). Basado en Jobling, 2004 tras adaptación de Sturm, 2001.

En esta ruta es regulada la biogénesis del melanosoma y es restringido su transporte hacia la periferia de la célula (Schiaffino-Tacchetti, 2005).

II.2.d.3. Pigmentación del iris

El estudio del color de ojos como rasgo físico en humanos, ha sido posible gracias al desarrollo de diversos estudios biológicos, morfológicos, químicos y genéticos que determinan la estructura del iris, el cual forma parte de la capa anterior del tracto uveal del ojo (Sturm, 2009).

El iris es una pequeña estructura muscular de tejido conectivo con una abertura central: la pupila, la cual esta situada detrás de la córnea y frente al cristalino, separando las cavidades anterior y posterior del ojo (Fig 6i).

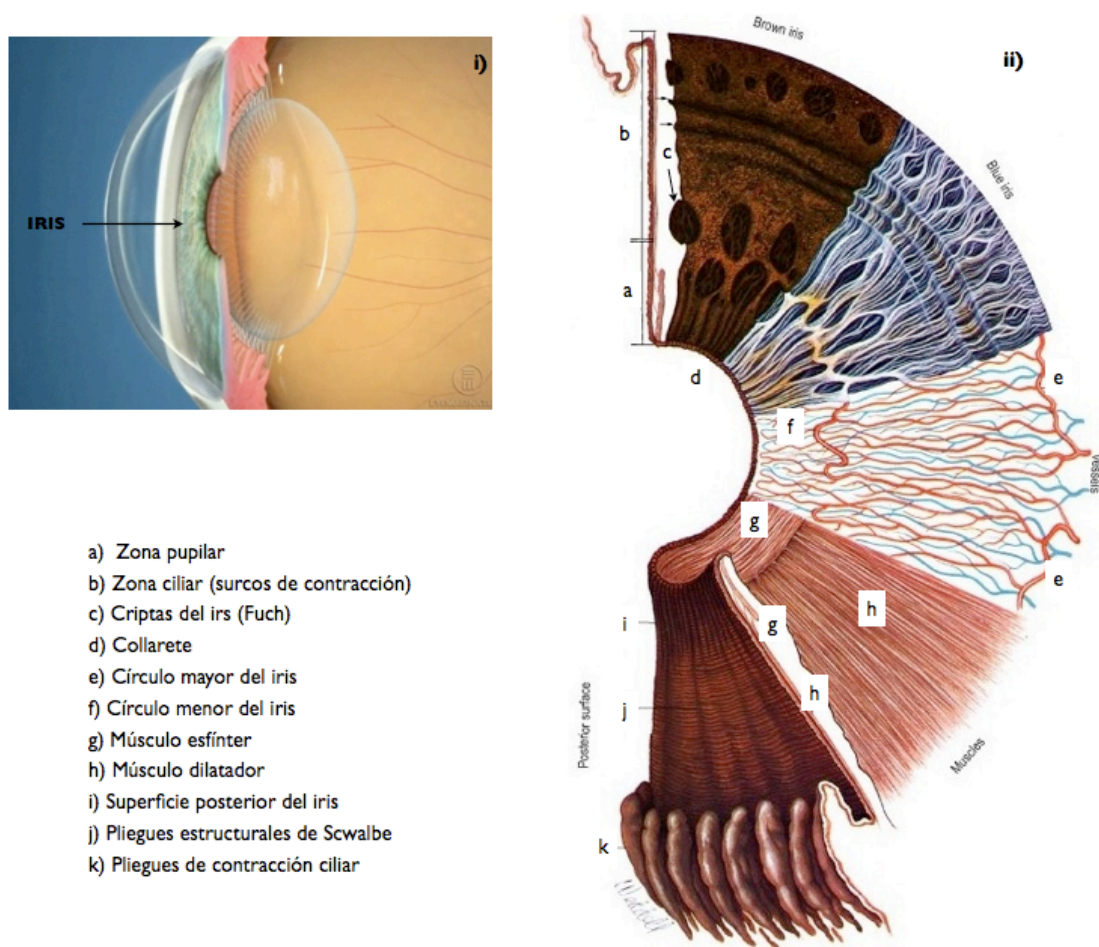


Figura 6. Ubicación (i) y anatomía del iris (ii). Fuente: (Gold-Lewis, 2004)

La función del iris radica en la regulación de la cantidad de luz que entra al ojo. Esta estructura comprende dos capas de tejido: la capa interior, la cual consta de células en forma de cubos que contienen pigmentos y que se encuentran parcialmente fusionadas, conocida como el epitelio pigmentario del iris (IPE). La capa más externa, conocida como capa anterior del estroma está compuesta por diversos arreglos de tejido conectivo, fibroblastos, y melanocitos. La presencia de los pigmentos de melanina en el iris es uno de los responsables de la impresión visual de la coloración del ojo en humanos. Las células del melanocito son agregadas hacia el borde anterior del estroma iridial, de forma paralela a la superficie del ojo (Jones et al., 1996; Sturm-Frudakis, 2004). Estos melanocitos, normalmente alcanzan la cantidad de melanina determinada genéticamente en la infancia temprana, y usualmente permanecen constante en la edad adulta (Bito et al., 1997). Además del IPE y los melanocitos, existe una tercera categoría de células que contienen melanina, conocidas como “células agrupadas” o *clump cells*, similares a macrófagos cargados de pigmentos. Estas células agrupadas se encuentran en el estroma y actúan fagocitando los gránulos de melanina y los melanocitos perdidos, es decir, aquellos que no fueron agregados. Éstas también se han descrito como células que han migrado desde el epitelio pigmentario del iris (Imesch et al., 1997)

Aparte de los individuos albinos, los cuales carecen de pigmentos de melanina y cuya apariencia visual en los ojos puede presentar una coloración rosa (producto de una reflexión de la luz por parte de las cavidades sanguíneas (Fig. 6ii-e), el IPE no ejerce una mayor influencia en la percepción de la coloración normal del ojo (es decir, en aquellos individuos sin alteraciones en este rasgo), debido a que la melanina presente en esta capa se encuentra distribuida de manera similar entre individuos que posean diferentes colores de iris. Notablemente, son la densidad y composición celular en los melanocitos del estroma del iris quienes deben ser considerados como los principales factores que afectan la coloración del ojo (Imesch et al., 1997; Sturm-Frudakis, 2004).

Los melanocitos en el estroma del iris con altos niveles de melanina absorben mayor cantidad de luz, dando la apariencia de coloración oscura. Cuando existe una ausencia de melanina en la capa anterior del ojo, la luz entra a través del estroma y es

dispersada por el colágeno el cual absorbe la mayoría de los colores exceptuando el azul y el gris los cuales, por consiguiente, son reflejados. Los arreglos de colágeno más finos tienden a dar una apariencia más azul, mientras que aquellos más gruesos dan la apariencia de coloración gris. Las tonalidades intermedias como verdes y avellanas, son el resultado de una variedad en las cantidades de melanina que permiten que la luz entre a través del estroma, reflejando una mezcla de tonalidades y sombras (Imesch et al., 1997). Por otra parte, la propia densidad del iris puede determinar la presencia de tonalidades en las coloraciones intermedias (Sturm-Larsson, 2009). También se ha descrito la presencia de un anillo peripupilar de pigmentación oscura en el iris asociado a las tonalidades intermedias, aunque el conocimiento en este aspecto ha sido limitado (Sturm-Frudakis, 2004). Entre los principales patrones complejos observados en el iris que contribuyen a la impresión visual de la coloración del ojo, están las criptas de Fuch (Fig. 6ii-c), los surcos de contracción (Fig. 6ii-b) y las acumulaciones de colágeno, los cuales comparten algunos factores genéticos con otras características del iris que pueden igualmente afectar el aspecto de su coloración aunque de forma indirecta (Sturm-Larsson, 2009).

Por todos estos factores, la asignación de fenotipos para algunas tonalidades de este rasgo resulta difícil de realizar objetivamente (Frudakis et al., 2007). Sin embargo, debe reconocerse también que clasificar el color de ojos únicamente como azul, intermedio, y marrón puede resultar simple, debido a la existencia de un rango continuo de coloraciones, tal y como las observadas especialmente entre individuos europeos (Sturm-Frudakis, 2004). En este sentido, se han desarrollado diversos métodos de clasificación, basados en mediciones cuantitativas, valorando parámetros como la saturación y tonalidad, mediante el empleo de fotografías digitales (Liu et al., 2010) y programas informáticos que analizan la luz reflejada por el iris, donde la suma entre la luminosidad y la escala de color permite realizar estimaciones sobre el índice de melanina del iris (IMI) (Frudakis, 2008). También se han descrito metodologías de clasificaciones cualitativas, empleando cartillas de colores (Valenzuela et al., 2010), así como también una variedad de términos descriptivos ya sea por autoevaluación, por el uso de un observador entrenado para realizar la puntuación, o empleando fotografías estándar para la clasificación (Franssen et al., 2006; Seddon et al., 1990). Estos métodos pueden ser considerados como subjetivos, y no son totalmente fiables como aquellos en donde se emplean registros de fotografías

automatizadas, en las cuales se ha procurado una mejora para la determinación del gradiente de coloraciones en este rasgo (Takamoto et al., 2001; Niggemann et al., 2003). Por su parte, existen otros patrones complejos que presenta el iris aparte de los que pueden afectar la apariencia de su coloración, los cuales comúnmente son empleados como biomarcadores para diversos propósitos como por ejemplo, la identificación individual por lectura de reconocimiento del iris. Estos análisis biométricos han demostrado que el iris presenta una complejidad que abarca aproximadamente 240 grados de libertad (Daugman, 2003).

II.2.d.4. Investigación genética del color de ojos

La expansión del conocimiento sobre los genes que están involucrados en la pigmentación humana, ha sido posible gracias a la combinación de diversos enfoques genéticos, bioquímicos y celulares, así como por la disponibilidad al acceso de la secuencia completa del genoma humano, y la gran información existente sobre SNPs entre las diversas poblaciones humanas (Sturm, 2009).

Uno de los locus responsables del fenotipo azul/no azul fue identificado por estudios de ligamiento en el cromosoma 15q por Eiberg y Mohr en 1996 (Eiberg-Mohr, 1996) (OMIM 227220). A través de mapas de ligamiento en el locus candidato 15q12-13, fue posible establecer la región flanqueante por los marcadores D15S165 y D15S144, con un valor máximo de LOD *score* cercano a D15S165, posteriormente se sugirió *OCA2* como gen candidato para este rasgo (Eiberg et al., 2008).

- *OCA2*: Es el homólogo del gen de los ojos rosa en ratones (P). En humanos, el gen *OCA2* se divide en 24 exones que cubren 345 Kb, de estos, 23 exones abarcan 836 regiones codificantes aminoácidas, y su exón 1 representa exclusivamente la región no codificante 5' UTR. Este gen produce una variación en la proteína integral de membrana P, la cual contiene 12 dominios transmembrana, que ayudan a regular el proceso de la melanogénesis. El conocimiento sobre la función específica de la proteína *OCA2* es ambiguo y se ha sugerido como un transportador anitporte de Na⁺/H⁺ (Eiberg et al., 2008), o como un transportador de glutamato (Lamoreux et al., 1995); ambas funciones indican que la proteína *OCA2* puede estar involucrada en el tráfico

intracelular de la enzima tirosinasa, regulando el pH durante la maduración del melanosoma (Toyofuku et al., 2002). Se han propuesto estudiar la variabilidad de diversos SNPs dentro de este gen empleando análisis de diplotipos (combinación específica de dos haplotipos) para una futura predicción de tipo más exacta sobre el contenido de melanina en el iris a partir del ADN (Duffy et al., 2007; Eiberg et al., 2008). Así como también se ha demostrado que las asociaciones de las variables contenidas en este gen al color de ojos, son independientes del grupo ancestral biogeográfico que sea estudiado (Frudakis et al., 2003b; Frudakis et al., 2007). Dentro de este gen, una de las mayores asociaciones encontradas al color de ojos azul frente a no azul, ha sido observada en los SNP rs7495174, rs4778241, rs4778138 (formando un bloque haplotípico en el intrón 1 de este gen) (Sturm et al., 2008) y rs1375164 (en el intrón 2) (Duffy et al., 2007). Por otra parte, el SNP rs1800407 presenta una fuerte asociación con el color de ojos no-azul (Duffy et al., 2007). Otros de las asociaciones detectadas al color de ojos ha sido observada en rs4778232 y rs8024968 (Kayser et al., 2008).

- HERC2: Este gen localizado en el sentido 3' UTR del gen *OCA2*, codifica los dominios HECT y RCC1 involucrados en el tráfico de proteínas (Kayser et al., 2008; Sturm et al., 2008). Tanto en éste gen como en *OCA2* se encuentran contenidos los SNPs con la mayor asociación al color de ojos. Se identificaron diversos SNPs relevantes en este gen, mediante análisis de mapas de asociación (Sturm et al., 2008), y por GWAS (Kayser et al., 2008). El SNP rs12913832 está situado en la región intergénica en el sentido 3'UTR a 21.152 bases respecto al gen *OCA2*, formando parte del gen *HERC2*. Este SNP se ubica en el centro de una secuencia corta y muy conservada entre diversas especies animales, por lo que ha sido ampliamente estudiado. En dicha secuencia, la variable asociada a ojos azules/marrones aparece con una frecuencia del 78%. La secuencia circundante a esta variable forma un sitio consenso de enlace para la familia de los factores de transcripción de la helicasa (HLTF *helicase like transcription factor*) (Sturm et al., 2008). Se ha descrito que este SNP es capaz de predecir con mayor exactitud el color de ojos azul/no azul que los haplotipos estudiados en *OCA2* (Sturm et al., 2008). Sin embargo, este único SNP en el ADN no puede explicar todas las

variaciones del color marrón, tomando en cuenta el rango existente desde el color avellana hasta los ojos con manchas marrones de melanina. El SNP codificante rs1800407 en *OCA2*, actúa con cierta penetrancia como un modificador sobre el SNP en *HERC2* para el color de ojos, y de alguna forma independientemente del riesgo de melanoma (Eiberg et al., 2008). Por su parte, el SNP rs1129038 se localiza en la región 3'UTR del exón 93 del gen, a 12,4 Kb del primer exón de *OCA2*. Tanto rs12913832 y rs1129038 han sido reportados en diversos estudios como fuertemente asociados al color de ojos azul, y ambos SNPs fueron encontrados en posición *cis*, evidencias que permitieron sugerir a estos SNPs como probables variables causales o que se encuentran en fuerte LD con la variable causal (Sturm et al., 2008). Sin embargo, de acuerdo al estudio de Sturm en 2008, se encontró que el SNP rs1129038 predice un poco mejor el color de ojos incluso que rs12913832, aunque el rs12913832 se encuentra contenido en una secuencia conservada. En un estudio funcional, Eiberg sugirió que estos dos SNPs poseen diferentes actividades reguladoras del gen *in vivo*, de acuerdo a experimentación realizada en líneas celulares Caco2, en las cuales éstos alelos demostraron efectos reguladores distintos sobre la actividad promotora de *OCA2*, apoyado también por estudios en EMSA, en el cual fueron encontrados diferencias en la movilidad electroforética de unión de los factores nucleares de estas células Caco2 a los alelos de estos marcadores asociados al color azul y marrón (Eiberg et al., 2008). Sin embargo, en contraste con este modelo basado en la activación, se ha propuesto un modelo de represión de la transcripción. La acción represiva observada a través de ensayos de transfección en la región conservada del intrón 86 en el gen *HERC2*, fue empleada para extrapolar la función silenciadora transcripcional de este elemento en melanocitos evaluando la expresión del gen *OCA2*. De acuerdo con este modelo, se desencadena una cascada de interacciones moleculares, en donde dependiendo del alelo presente en el SNP rs12913832, podrá o no ser inducido el desdoblamiento del estado inicial de empaquetamiento de la región *OCA2-HERC2*, dando lugar a un estado relajado de eucromatina, la cual permite que la región promotora de *OCA2* esté disponible para la transcripción. Una vez desenrollada esta región, el factor de transcripción HLTF actúa como un miembro del complejo de remodelación SWI-SNF,

reconociendo la secuencia específica de ADN contenida en el intrón 86 del gen *HERC2*: rs12913832*T. Cuando esta unión ocurre, se enlazan también a este *locus* de la región control los factores de transcripción MIFT y LEF1. Este evento promueve un cambio que ocurre a 21 kb del promotor inmediato de *OCA2* en sentido 3' UTR, el cual permite que el complejo Po II (RNA Polimerasa II) inicie la transcripción. El producto de esta transcripción es la proteína *OCA2* la cual estimula la maduración de los melanosomas y los niveles altos de melanocitos en el iris dando como resultado el color de ojos marrón (Fig. 7a). Cuando el alelo rs12913832*C está presente en el intrón 86 del gen *HERC2*, el cambio en esta base previene la interacción del HLTF con la heterocromatina, el fallo en la unión de MIFT y LEF1, y por consiguiente, el promotor de *OCA2* permanece cerrado. La ausencia de la proteína *OCA2* en los melanocitos del iris, dan como resultados melanosomas inmaduros y una pérdida en la producción de melanina originando la presencia de ojos azules (Fig. 7b). Dado la importancia de gen *HERC2* en la predicción del color de ojos, se han realizado algunos estudios que han explorado la variabilidad de otros SNPs contenidos en esta región (Sturm et al., 2008), e inclusive combinando haplotipos de este gen con otros descritos en *OCA2* (Mengel-From et al., 2010). De acuerdo a este mecanismo molecular propuesto por Sturm y Larson, se han observado incrementos significativos en los niveles de transcripción de mRNA de *OCA2* para rs12913832*T (marrón) comparado con el alelo rs12913832*C (azul) en secuencias de melanocitos humanos (Cook et al., 2009), lo cual es consistente con el modelo. Recientemente, también se evaluó la afinidad de los factores de transcripción HLTF, LEF1, y MITF al rs12913832*T observando la formación de un *loop* o lazo formado entre esta secuencia y el promotor de *OCA2*, la cual permite un aumento en la transcripción de este gen. En contraste, cuando el alelo rs12923832*C está presente, se ha observado una reducción tanto del lazo de cromatina, como de la unión de los factores de transcripción, y por lo tanto de la producción de *OCA2*, demostrando no solo que esta variación alélica interrumpe el potencial regulador de este elemento, sino que se afecta su interacción con el promotor en cultivos celulares (Visser et al., 2012).

- Otro SNP reportado en *HERC2* con una fuerte asociación al color de ojos además de rs12913832 y rs1129038 es rs1667394, el cual ha mostrado una correlación de azul frente a marrón y azul frente a verde en población de Islandia (Sulem et al., 2007), y cuya asociación con el color verde se ha confirmado también en una cohorte de *23andMe* (Eriksson et al., 2010). Recientemente, en un estudio funcional se ha encontrado que la variable del SNP rs7183877 incrementa la transcripción de *OCA2* de manera similar a rs12913832, aunque debido a la resolución de la técnica empleada, se ha atribuido dicha observación a la proximidad existente entre estos marcadores (Visser et al., 2012). Por otra parte, se ha descrito una asociación de este marcador (rs7183877) con el color de ojos verdes (Eriksson et al., 2010). Los SNP rs916977, rs11636232 han sido descritos también como relevantes en análisis de asociación al color de ojos, aunque estos se encuentran en fuerte LD con rs12913832 (Kayser et al., 2008; Mengel-From et al., 2010).

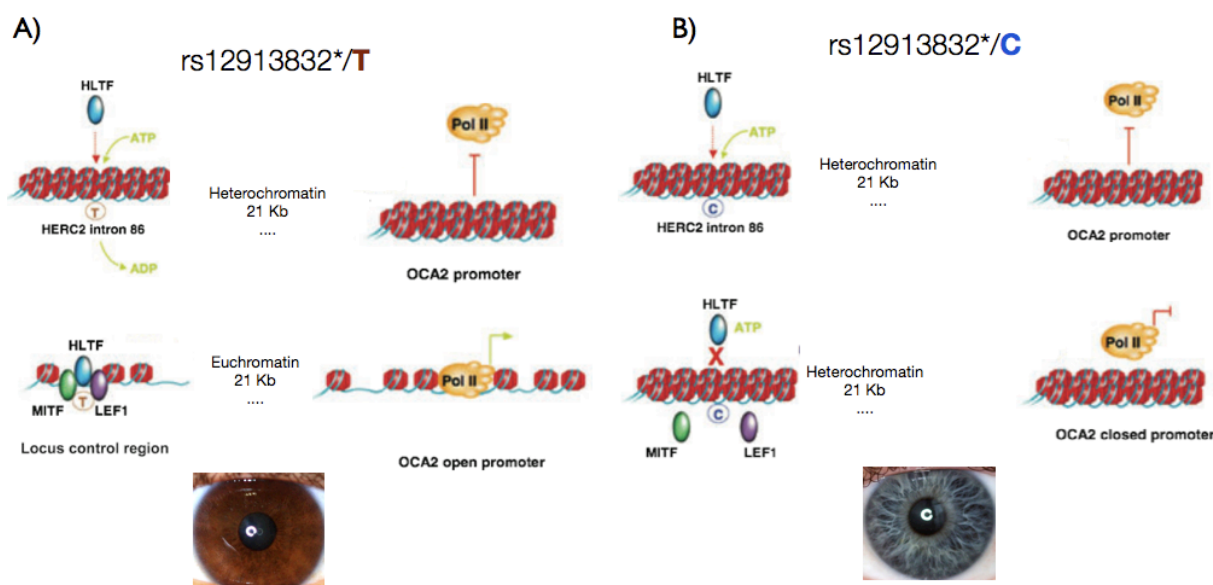


Figura 7. Modelo propuesto por Sturm & Larson en 2009 para la determinación del color de ojos azul/marrón por regulación de la expresión del gen *OCA2*. Fuente: basado en Sturm y Larson, 2009.

Uno de los primeros estudios donde fue evaluado la asociación de otros genes aparte de *OCA2*, fue realizado por T. Frudakis en el 2003. En este trabajo, se realizó una exploración sobre algunos genes candidatos relacionados con las rutas bioquímicas de la pigmentación, como *ASIP*, *TYR*, *TYRP1*, *DCT*, *MC1R*, entre otros, encontrando una asociación significativa al color de ojos para estos genes a nivel haplotípico, la cual explica un 13% de la variabilidad encontrada en la población de

estudio (caucásica). Por otra parte, se ha determinado que alguna de estas secuencias forman diplotipos que explican un 15% de la variación, mientras que sólo se ha encontrado una menor variabilidad explicada por los SNPs analizados individualmente en estos genes. Dichos hallazgos han sugerido la presencia de un elemento de complejidad intragénica en la determinación del color de ojos, como puede ser la debida a interacciones de tipo epistáticas. Se debate además si estas asociaciones menos fuertes encontradas en estos genes, se deben a una baja tasa de penetrancia alélica, o a la presencia de secuencias informativas para ciertas poblaciones con sub-estructuras crípticas que guardan una relación con un determinado color de ojos (Frudakis et al., 2003a; Sturm-Frudakis, 2004).

En la actualidad, mediante diversos estudios de asociación de genomas completos, se ha confirmado la relación de algunos de estos haplotipos con el color de ojos, pero también ha sido posible el conocimiento sobre algunos SNPs específicos que muestran una asociación individual, los cuales están contenidos en algunos de estos genes de efectos menores como *TYRP1* y *TYR*, así como en otros genes los cuales no habían sido identificados en estudios previos, como (*MATP*)*SLC45A2*, *SLC24A4* y *IRF4*, quienes además mostraron estar asociados al color de cabello y piel. Adicionalmente, en estos estudios se ha explorado la posible asociación con otros genes como *MC1R*, *TPCN2* y *ASIP*, dada su fuerte correlación con pigmentación en piel y cabello, encontrando una asociación con el color de ojos muy débil, o no detectada entre las poblaciones estudiadas.

- *TYRP1*: En este gen, el SNP rs1408799 presenta una asociación significativa cuando se compara el color de ojos azules frente a no azules en población de Islandia y Holanda (Sulem et al., 2008). El haplotipo formado por este SNP junto con rs683, de igual forma confirma dicha asociación bajo este mismo análisis realizado también en población de Polonia (Pospiech et al., 2011).
- *TYR*: los SNPs encontrados en este gen rs1042602 y rs1393350, aunque presentes en el mismo bloque de LD en población de Islandia, sus efectos sobre la pigmentación son diferentes. El SNP rs1393350 presenta una asociación que se aproxima a la significancia de genomas completos, cuando se comparan clases de colores de ojos azules frente a verdes, mientras que

rs1042602, aunque previamente reportado como asociado al color de ojos, de acuerdo al estudio de Frudakis, también se ha encontrado asociado con la presencia de pecas en población de Islandia (Sulem et al., 2007). Esta doble asociación también fue observada en el SNP rs1126809, el cual aunque también asociado al color de ojos, presenta una correlación con la presencia de sensibilidad de la piel al sol (Sulem et al., 2008).

- MATP (SLC45A2): Los SNPs, rs6867641, rs13289 han mostrado una asociación con la variación de coloración de piel evaluado en poblaciones Caucásicas, Africanas, Asiáticas y Nativos Australianos. Por su parte, rs16891982 el cual se encuentra en la región promotora de este gen, ha mostrado una asociación con el color de cabello y piel, así como una asociación secundaria al color de ojos marrón (Graf et al., 2005). Se ha descrito que rs16891982 es el marcador de este gen con mayor probabilidad de ser una variable causal o de estar en fuerte desequilibrio de ligamiento con la variable causal en *MATP*. Las frecuencias alélicas encontradas entre los cuatro grupos geográficos antes mencionados, fueron significativas para los SNPs rs26722, rs13289, rs6867641 y rs16891982 (Graf et al., 2007).
- SLC24A4: El SNP rs12896399 ha mostrado una asociación significativa tanto para la coloración ojos azul frente a verde así como en color de cabello rubio frente a marrón, en poblaciones de Islandia y Holanda (Sulem et al., 2007). En otro estudio, se ha confirmado la asociación de este SNP al color de cabello claro, y una correlación secundaria con la tendencia de bronceado en piel. Tanto este marcador como algunos en *SLC45A2* explican aproximadamente un 21,9% de la variación residual en del color de cabello (rubio-negro)(Han et al., 2008).
- IRF4: El SNP rs1540771 ha mostrado una asociación con la presencia de pecas en población de Islandia, así como asociaciones secundarias con cabello marrón y sensibilidad de la piel a los UVR. Se ha encontrado además que el alelo A de este SNP es más frecuente en población europea, lo cual sugiere que posiblemente exista una selección positiva debida a los efectos en la reducción en pigmentación de piel de este gen. Adicionalmente, se ha

sugerido que este SNP posee una fuerte correlación con ciertas zonas geográficas de Inglaterra (Sulem et al., 2007). Por su parte, el SNP rs12203592 ha mostrado una asociación al color de ojos tras un análisis de genomas completos en una población holandesa (Liu et al., 2009), y específicamente a color de ojos claros en otro estudio realizado en individuos con ascendencia caucásica en Estados Unidos y Australia, así como también ha mostrado una fuerte asociación al color de cabello, cuando se comparan los grupos negro-rojo, y rubio-negro, explicando un 7% de la variabilidad encontrada en este rasgo, y en menor medida también se ha encontrado asociación al color de piel (Han et al., 2008).

- ASIP: se han descrito haplotipos conformados por SNPs como rs1015362 y rs4911414, que están significativamente asociados al color de cabello rojo, la presencia de pecas, y sensibilidad de la piel al sol. La correlación de rs6058017 con la pigmentación oscura en piel ha sido observada en población africana, sin embargo, esta asociación no ha podido ser replicada en población asiática (Stokowski et al., 2007), ni en caucásica (Sulem et al., 2008). El análisis de este marcador junto con rs1800404 en *OCA2* ha sugerido un posible patrón de divergencia entre poblaciones de África del este y el resto (Norton et al., 2007).
- MC1R: Se ha descrito que dos variantes no sinónimas en este gen comunes en población europea del CEPH, las cuales tienen un efecto mayor en la pigmentación de esta población: rs1805007 y rs1805008, mientras que dicha variabilidad no ha sido observada en poblaciones del este asiático (ASN), ni en Nigerianos Yoruba (YRI), lo cual sugiere que estas variaciones han estado levemente afectadas por una selección positiva reciente (Sulem et al., 2007).
- TPCN2: Se ha descrito que los SNPs rs35264875 y rs3829241, los cuales se encuentran en el mismo bloque de LD, han mostrado una asociación significativa cuando se comparan los colores de cabello rubio frente a marrón en población de Islandia (Sulem et al., 2008).

Dada la complejidad que presentan los mecanismos de regulación genética en la pigmentación humana, se han comenzado investigaciones sobre las posibles interacciones genéticas que puedan explicar parte de dicha variabilidad. En el año 2004, se realizó un estudio en población caucásica, en donde se detectó un efecto interactivo entre *MC1R* y *OCA2* que afecta el color de piel y la presencia de pecas (Duffy et al., 2004). Posteriormente en el 2008, un estudio de genomas completos indicó la interacción existente en SNPs contenidos dentro del gen *OCA2*, así como entre SNPs de este gen y de *HERC2*, atribuyendo estas interacciones en parte a la existencia de haplotipos (Kayser et al., 2008). En el 2009, se demostró la existencia de interacción entre los genes *MC1R* y *HERC2* que afectan la determinación de color de piel y cabello, específicamente describiendo que el alelo rs12913832*T en *HERC2* podría ser epistático sobre la mayoría de las funciones de los alelos de *MC1R* (Branicki et al., 2009). En un estudio más reciente realizado sobre interacciones genéticas en la variabilidad del color de ojos, se determinó la existencia de interacciones redundantes (en donde la interacción provee información redundante, basado en una entropía negativa) entre *HERC2* y *OCA2* que afectan al color de ojos avellana, así como entre *HERC2* y *SLC24A4* afectando el color de ojos azules. En este estudio, también se indicó que existen efectos interactivos pero de carácter sinérgico (en donde las interacciones entre los genes ofrecen más información que la suma del efecto de los genes individuales, basado en una entropía positiva) entre *HERC2* y *OCA2*, así como entre *HERC2* y *TYRP* ambos afectando la determinación del color de ojos verdes (Branicki et al., 2011).

Uno de los modelos de predicción de color de ojos propuestos a partir de los conocimientos sobre los determinantes genéticos existentes, ha sido presentado en el 2010 y se conoce como *Irisplex*. La selección de los marcadores empleados en este modelo, está basado en los datos de un análisis previo de genomas completos, realizado en Rotterdam (Holanda)(Kayser et al., 2008). Tras evaluar modelos de árboles de clasificación, redes nodulares, agrupamiento *Fuzzy-C*, y regresión ordinal, encontraron una clasificación más exacta de las clases fenotípicas, definidas como azul, intermedio y marrón cuando se aplica el modelo de regresión, con valores de AUC de 0,91; 0,70 y 0,93 respectivamente, los cuales fueron estimados para los 24 SNPs que encontraron asociados en esta población. Sin embargo, los autores

atribuyeron el mayor peso en el poder de predicción, sólo a los seis primeros marcadores establecidos de acuerdo al *ranking* de la regresión logística multinominal, que se encuentran en seis genes diferentes (*HERC2*, *OCA2*, *SLC45A2*, *SLC24A4*, *IRF4* y *TYR*) (Liu et al., 2009), determinando además que la predicción del color de ojos del *Irisplex* es independiente de la información sobre el ancestro biogeográfico (Walsh et al., 2011b). Posteriormente, este test fue validado para la investigación forense siguiendo las recomendaciones del grupo de trabajo científico en métodos de análisis de ADN (SWGAM *Scientific Working Group on DNA Analysis Methods*) (Walsh et al., 2011a), así como también fue evaluado en siete poblaciones de Europa, donde se demostró empíricamente que este modelo ofrece una predicción más exacta especialmente de los colores de ojos azules y marrones, mientras que para el color de ojos intermedios entre los que se incluye el verde, los autores sugieren la realización de futuros estudios que permitan una mejora en la predicción de estas tonalidades intermedias (Walsh et al., 2011c). Justamente, la limitación que posee el modelo del *Irisplex* radica en una ausencia de sensibilidad (definida como el porcentaje de acierto en la clasificación) en la predicción de fenotipos intermedios, ya que el valor de AUC para este grupo en realidad se atribuye sólo a una alta especificidad (definida como el porcentaje de acierto en el descarte). Estos fallos en la clasificación de fenotipos intermedios se han observado también en diversos estudios realizados en poblaciones de Alemania (Purps et al., 2011), Turquía (Bubul et al., 2011), y población con mezcla Euro asiática de Australia (Prestes et al., 2011).

II.2.e. Consideraciones éticas y legales sobre la inferencia de EVCs a partir del ADN

En el 2009, M. Kayser y P. Schneider realizaron una revisión sobre los aspectos éticos y legales en el uso de marcadores genéticos para la inferencia de EVCs (Kayser-Schneider, 2009) posteriormente actualizado por Kayser y Knijff en 2011 (Kayser-de Knijff, 2011). La inferencia de EVCs a partir de marcadores genéticos está permitido y regulado por la ley únicamente en Holanda. En ningún otro país se ha introducido una ley específica que pueda permitir el empleo de este tipo de evidencia en un caso forense. Sin embargo, el servicio de ciencias forenses del Reino Unido y los Estados Unidos, ha empleado algunas predicciones de EVCs basadas en análisis del ADN justificadas por leyes ya existentes. Los sistemas civiles legales de los países europeos

(apartando Holanda) sólo permite la investigación sobre una base legal específica, que de hecho prohíbe la inferencia de FDPs, al menos que específicamente esté provista por la legislación. En algunas jurisdicciones de USA sólo prohíben la determinación de enfermedades genéticas. Además de las implicaciones legales, el uso de marcadores genéticos para la predicción de EVCs representa un desafío desde el punto de vista ético, y requiere ser un tema a discutir abiertamente tanto por la comunidad científica así como por la sociedad en general.

Una de las principales distinciones realizadas en esta revisión, se refiere al uso de marcadores en el ADN con propósito forense que sean “codificantes” frente a “no-codificantes”. Aunque ésta parezca ser una definición que delimita estrictamente la porción del genoma que puede ser analizado, debe tomarse en consideración que nuestro genoma está organizado en bloques de ADN que son heredados mayormente intactos. Por lo tanto, un marcador que se encuentre en la región no codificante y que esté físicamente próximo al marcador que codifica para un fenotipo dado, puede revelar la misma información que el marcador codificante, debido al desequilibrio de ligamiento. La mayoría de los marcadores genéticos asociados a EVCs se encuentran en regiones no-codificantes, pero están genéticamente ligados a la variante causal del efecto funcional. En consecuencia, el empleo de estos marcadores no debería violar la legislación si ésta sólo permite que se apliquen los marcadores no-codificantes a investigaciones forenses. Por otra parte, la apariencia física de un individuo es visible ante cualquiera, por lo tanto, la información que revela la predicción genética de los EVCs, no debería ser considerada como una información “privada”. La inferencia estadística obtenida a partir de los marcadores genéticos puede ser empleada para reducir el número de sospechosos bajo el mismo principio de un reporte realizado por un testigo ocular, basado en el grado de verosimilitud de la información aportada. Finalmente, los autores sugieren que el empleo de marcadores genéticos para inferir EVCs no debe sesgarse en contra de ninguna población, por el contrario, esta inferencia provee una predicción sustentada en la verosimilitud incluso con mayor precisión que la ofrecida por un testigo ocular, dependiendo del EVC y de la habilidad de comprensión sobre la complejidad biológica que sea analizado.

II.3. Algunos casos aplicados al estudio de AIMs

II.3.a. Operación Minstead

En el año 2008, T. Frudakis dedicó un apartado sobre este caso en su libro “*Molecular Photofitting*”, luego de su participación junto a *DNAPrint Genomics*, con el propósito de aplicar el *test* de inferencia de grupos ancestrales llamado *DNAWitness™*.

Reseña del caso: La operación Minstead se estableció en 1998 para aprehender a un ladrón y violador el cual estuvo asechando a mujeres en el sur de Londres desde el año 1992. En el año 2004 se documentaron más de 80 agresiones - la mayoría de ellas en mujeres de la tercera edad, entre 68 y 93 años- y el caso llegó a ser por mucho, la investigación de un violador en serie más grande que haya realizado la policía Metropolitana de Londres, en términos de número de víctimas, número de sospechosos, y tiempo implicado en la investigación. Las evidencias obtenidas a partir del ADN en los casos indicaron que se trata de un mismo agresor. Una vez dentro de las casas, el ladrón empleaba un procedimiento de operación estandarizada; primero desconectaba el teléfono y la electricidad, sacaba las bombillas de luz de las habitaciones, y eventualmente cuando la víctima regresaba de una caminata fuera de casa, él solía pedir dinero. Usualmente pasaba mucho tiempo en casa de las víctimas. Cuando ocurrió la primera agresión en 1992, este ya había realizado varios robos. Existieron tres fuentes de información sobre el agresor: La primera procedía de un patrón cronológico inusual en donde ocurrieron las agresiones. La segunda, se obtuvo a través de las víctimas, las cuales habían descrito al agresor como de piel oscura, aproximadamente de 30 años, 1,80 metros de altura y de constitución atlética. Debido al hecho de que los crímenes ocurrieron durante la noche, en habitaciones sin iluminación, hubo una incertidumbre considerable sobre la apariencia del agresor ya que otras víctimas lo describieron de piel clara. La tercera fuente fue el ADN dejado en la escena del crimen. Los investigadores no fueron capaces de encontrar concordancias entre los perfiles de STRs en sus bases de datos, e incluso realizaron un muestreo entre cientos de individuos del área del sur de Londres sin poder encontrar ninguna concordancia. En marzo de 2004, *DNAPrint* utilizó el ensayo de 171 AIMs llamado *DNAWitness™* en el caso y obtuvieron un MLE (estimación de máximo de verosimilitud) de 82% Africano subsahariano, 6%


Europeo y 12% Nativo Americano. Las estimaciones posteriores de las proporciones de mezcla revelaron que uno de los padres tenía considerablemente un grupo ancestral más subsahariano que el otro, indicando además que la mezcla del individuo ocurrió recientemente. La información conocida sobre el caso hasta la fecha era: (a) la contribución de su grupo ancestral no africano fue realizado principalmente por uno de sus padres. (b) su grupo ancestral (así como el del padre con mezcla) es más verosímil afro-caribeño en lugar de cualquier otra ascendencia africana presente en Gran Bretaña. (c) El padre contribuyente al grupo ancestral no africano, es más verosímil que provenga del área de las islas Windward (sureste) que de Jamaica (noroeste), pero no pueden descartarse otras islas occidentales. En enero de 2007, el caso aún seguía sin resolverse, pero las investigaciones sobre la lista de posibles sospechosos se habían reducido a 1000 hombres que concordaban con el perfil biogeográfico (Frudakis, 2008). Posteriormente, un segundo estudio realizado sobre este caso, reveló una variación en las proporciones de mezcla antes reportadas (85% AFR, 12 EUR, 3%AME) siendo posible que dicho patrón sea originario de cualquier población del Caribe o cercana, y no específicamente del sureste. Esto, sumado al bajo número de muestras de referencia que fueron recolectadas en estas islas durante el estudio, la ausencia de un error de clasificación asociado al *test* y a la detección de un componente asiático observado en estas islas debido a una sobreestimación en el diseño del *test*, fueron algunas de las observaciones más relevantes realizadas durante este análisis (C. Phillips, comunicación personal). En marzo del año 2011, fue anunciado a través de diversos medios de comunicación que la investigación había culminado, tras declarar a Delray Grant (nacido en Jamaica) culpable de varias agresiones sexuales y robos cometidos en el sur de Londres.

II.3.b. Ataque terrorista en Madrid 11-M

Phillips, y colaboradores en el 2009 publicaron un estudio sobre las muestras obtenidas del ataque terrorista ocurrido en 11 de marzo del 2004 en Madrid, en el cual se analizaron tanto los marcadores genéticos de rutina del ADNmt y el cromosoma Y, como un conjunto de AIMs autosómicos (34-plex)(Phillips et al., 2007), para determinar el grupo ancestral posible al que correspondían estas muestras.

Reseña del caso: El 11 de marzo de 2004, 10 dispositivos de explosión improvisados (IED) fueron detonados en cuatro trenes de pasajeros de la red de

cercanías de Madrid, en un ataque coordinado en donde hubo 191 personas muertas y 1.755 heridas, lo que representó uno de los ataques terroristas de mayor magnitud ocurrido en Europa. El estudio de ADN incluyó el análisis de perfiles de STRs de 226 muestras de referencia y más de 600 muestras obtenidas a partir de fragmentos que incluían restos de IEDs detonados, vehículos de los sospechosos, un IED no detonado encontrado en la estación de tren El Pozo, y artículos personales recogidos en los sitios del ensamblaje de las bombas, o en los domicilios empleados por los sospechosos. Por su parte, siete perfiles completos de STRs originarios de cinco artículos personales, junto con una huella en la mochila de un IED no detonado encontrado en El Pozo, no mostraron concordancias con los sospechosos detenidos al comienzo del juicio en febrero de 2007. Siguiendo una orden judicial, las siete muestras que no coincidieron fueron el foco de un análisis de genotipado especializado en un intento de asignar el grupo ancestral al que corresponden estas muestras. Se solicitó realizar una comparación específica entre posible origen europeo o norte africano, para poder diferenciar únicamente entre estos dos grupos. Los análisis del 34 plex revelaron que tres de las siete muestras poseen una mayor probabilidad de corresponder al norte de África que Europa, otras 3 presentaban un perfil indefinido, posiblemente correspondiente a una mezcla (*population admixture*), y una muestra presentó más probabilidad de ser europea. El análisis de los AIM-SNPs mediante el empleo de un clasificador Bayesiano permitió realizar una valoración en la habilidad de diferenciar europeos de norte- africanos por validaciones cruzadas del conjunto de entrenamiento empleadas en los cálculos de verosimilitud. Uno de los sospechosos cuya inferencia genética mediante el análisis de los AIM-SNP correspondía al norte de África, y cuya material genético fue obtenido a partir de un cepillo dental, fue posteriormente identificado como de origen Argelino (Phillips et al., 2009).



2. JUSTIFICACIÓN Y OBJETIVOS



III. Justificación

En genética forense, a raíz de las limitaciones encontradas en la investigación criminal cuando no existen sospechosos de haber cometido un delito, o cuando en una identificación individual no hay concordancias con los perfiles genéticos comparados en las bases de datos, ha empezado a surgir la necesidad de obtener información adicional más allá de la determinada por el género. Esto es ahora posible mediante la inferencia sobre características que permitan esbozar, al menos en parte, el aspecto físico de un individuo en cuestión, ya sea a través de métodos indirectos por el estudio del ancestro biogeográfico (BGA), o empleando métodos directos por la determinación de características externas visibles (EVCs). Ambas herramientas de inferencia, podrían representar potencialmente un complemento de gran valor en la investigación forense, ayudando a reducir el universo de sospechosos de haber cometido un delito, e incluso ser un complemento de valor agregado cuando estamos en presencia de un testigo ocular con un testimonio dudoso o confuso. Por otra parte, la inferencia del ancestro biogeográfico y de las características visibles externas, contribuirían también en gran medida a dirigir la investigación genética durante la identificación de víctimas en desastres naturales o en restos cadavéricos, ya que por lo general, en estos casos se trata de muestras en descomposición, y cuyo material genético se puede encontrar en estado de degradación, por lo que también existe la necesidad de desarrollar e implementar técnicas de detección que sean sensibles ante este material biológico. En este sentido, el desarrollo de ensayos basados en la detección de SNPs, representan una metodología que se adapta a este contexto, y en el caso particular de los EVCs, sólo se han encontrado asociaciones a estos rasgos físicos, en aquellas variantes genéticas representadas por polimorfismos de una sola base.

Dado que la investigación en genética forense requiere a su vez de los estudios realizados en genética de poblaciones, resulta de interés que cada vez sean analizados más grupos, y especialmente aquellos que se encuentran poco caracterizados, los cuales suelen estar escasamente representados entre las diversas bases de datos genéticas, como es el caso de las poblaciones del continente americano. Por otra parte, el estudio de estas poblaciones contribuiría no sólo a la estimación de la historia ancestral de sus individuos, sino al control de la estratificación de poblaciones, minimizando el riesgo de falsos positivos.

IV. Objetivos


Los objetivos planteados en el presente trabajo de investigación comprenden el estudio de *Forensic DNA Phenotyping* (FDP), a través de la caracterización de poblaciones con AIMs y la inferencia de la pigmentación humana como EVC, por lo que pueden separarse de acuerdo a estos temas en dos grupos de objetivos:

Sobre el estudio de AIMs:

1. Desarrollo de paneles de AIMs-SNPs para la inferencia del grupo ancestral biogeográfico (BGA) en poblaciones del continente americano.
2. Análisis de la variabilidad mitocondrial de poblaciones americanas, escasamente caracterizadas.
3. Evaluación del poder informativo de SNPs autosómicos de identificación individual en poblaciones americanas.

Sobre el estudio de EVCs :

4. Diseño y optimización de reacciones *multiplex* de SNPs, para la inferencia de la variabilidad en la pigmentación humana, mediante SNaPshot.
5. Aplicación del *multiplex* en la predicción de pigmentación humana, específicamente en la inferencia del color de ojos en diversas poblaciones geográficas.
6. Empleo de sistemas de predicción estadísticos que discriminen y además proporcionen una medida de error de los perfiles genotípicos que se encuentran asociados a la probabilidad de presentar un tipo de pigmentación.
5. Demostrar el potencial forense del *multiplex* de predicción de pigmentación de ojos, mediante ensayos en muestras en estado de degradación así como de baja concentración.



3. RESULTADOS Y DISCUSIÓN



V. Preámbulo

En este trabajo los resultados están presentados en dos bloques correspondientes a los objetivos previamente planteados, en los cuales están contenidos los diversos trabajos de investigación realizados. Al final de cada bloque se incluye un apartado de discusión y perspectivas futuras sobre los tópicos desarrollados.

Bloque 1. Sobre la caracterización genética de poblaciones nativas y con mezcla en América, mediante el estudio de AIMs autosómicos, marcadores de linajes y de identificación individual:

- V.1. *“Analysis of 52-plex markers in four Natives American populations from Venezuela”*.
- V.2. *“Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas”*.
- V.3. *“A melting pot of multi-continental mtDNA lineages in admixed Venezuelans”*.
- V.4. *“PIMA: A population indicative multiplex for the Americas”*.

Bloque 2. Sobre el diseño metodológico, modelo estadístico de predicción, potencial forense y aplicación de SNPs en el estudio de FDP:

- V.5. *“Further development of forensic eye colour predictive tests”*.
- V.6. *“A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type”*.
- V.7. *“A researcher’s guide to STRUCTURE software: applications, parameter settings and supporting software”*.
- V.8. *“Assessing the forensic potential of an eye colour predictive test in challenging DNA”*.

Bloque 1.

***V.1. Analysis of the SNPforID 52-plex
markers in four Natives American
populations from Venezuela***

Y. Ruiz, M.A. Chiurillo , L. Borjas, C. Phillips, M.V. Lareu, A. Carracedo.

(*Forensic Science International: Genetics*, 2012, doi:10.1016/j.fsigen.2012.02.007)



Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensic Population Genetics – Short Communication

Analysis of the SNPforID 52-plex markers in four Native American populations from Venezuela

Y. Ruiz^{a,*}, M.A. Chiurillo^b, L. Borjas^c, C. Phillips^a, M.V. Lareu^a, Á. Carracedo^{a,d}^a Forensic Genetics Unit, Institute of Forensic Sciences “Luis Concheiro”, University of Santiago de Compostela, Spain^b Molecular Genetics Laboratory “Dr. Jorge Yunis-Turbay”, Universidad Centroccidental “Lisandro Alvarado”, Barquisimeto, Venezuela^c Medical Genetic Unit, Medicine Faculty, University of Zulia, Maracaibo, Venezuela^d CIBERER, Genomic Medicine Group, University of Santiago de Compostela and Galician Foundation of Genomic Medicine, Spain

ARTICLE INFO

Article history:

Received 21 November 2011

Accepted 21 February 2012

Keywords:

Venezuela

Native American populations

52-Plex assay

Autosomal SNPs

SNPforID

Human identification

ABSTRACT

The SNPforID 52-plex single nucleotide polymorphisms (SNPs) were analyzed in four native Venezuelan populations: Bari, Pemon, Panare and Warao. None of the population-locus combinations showed significant departure from Hardy–Weinberg equilibrium. Calculation of forensic and statistical parameters showed lower values of genetic diversity in comparison with African and European populations, as well as other, admixed populations of neighboring regions of Caribbean, Central and South America. Significant levels of divergence were observed between the four Native Venezuelan populations as well as with other previously studied populations. Analysis of the 52-plex SNP loci with *Structure* provided an optimum number of population clusters of three, corresponding to Africans, Europeans and Native Americans. Analysis of admixed populations indicated a range of membership proportions for ancestral populations consisting of Native American, African and European components. The genetic differences observed in the Native American groups suggested by the 52 SNPs typed in our study are in agreement with current knowledge of the demographic history of the Americas.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Populations

The populations included in this study represent four endogenous groups located in Venezuela, in the mid northern region of South America: Bari, Pemon, Panare and Warao. There is evidence to indicate early peopling of this territory ~13,000 years ago [1]. The more recent demography of South American populations shows a strong effect from the arrival of European colonizers and the African slave trade, since the 15th century [2]. Today each of the four populations studied are part of a protected group declared as ‘live patrimony’ in Venezuelan law [3], with the aim of protecting the cultural heritage of indigenous peoples. Studying the genetic variability in these populations is therefore of interest for understanding the structure and composition of native populations from this part of South America. Furthermore, equivalent studies in contemporary American populations, notably those with admixed origin (widely referred to as Mestizos but in this study meaning mixed Native American and European ancestry), are important for understanding forensic, anthropological and clinical implications of population variability because the levels of admixture are higher

than other parts of the world. Additionally, in the forensic field, the use of the SNPforID 52-plex is a viable and informative alternative when STR results are not definitive enough in complex relationship tests [4], as well as for typing severely degraded DNA [5].

The characteristics of the four native populations included in the study were as follows:

- Bari (37 individuals): Ethno-linguistic family: *Chibcha*. Regional distribution: Serranía de los Motilones, Catatumbo Basin (border between Venezuela and Colombia). Sampling location: Machiques de Perijá (10°042'N 72°342'E), Zulia state, Venezuela.
- Pemon (34 individuals): Ethno-linguistic family: *Caribe*. Regional distribution: Bolívar state (Venezuela), NE Guyana, Northern Brazil. Sampling location: Santa María de Wonken (5.1167°N, –61.7667°W), Bolívar state, Venezuela.
- Panare (44 individuals): Ethno-linguistic family: *Caribe*. Regional distribution: west of Bolívar state and north of Amazonas state (Venezuela). Sampling location: Maniapure (6.9217°N, –66.5419°W), Bolívar state, Venezuela.
- Warao (24 individuals): Isolated linguistic group. Regional distribution: Delta Amacuro, East of Sucre and Monagas state (Venezuela), Surinam, and North Guyana. Sampling location: Islas tres caños (8.6425°N, –61.9847°W), Delta Amacuro state, Venezuela.

* Corresponding author.

E-mail address: yarimarruiz@gmail.com (Y. Ruiz).

A map showing the sampling location for these population groups is available in [Supplementary data Fig. S1](#).

We performed a comparative analysis with genetic data previously collected from populations of Ecuador (42 individuals), Europe (1140 individuals) and Africa (512 individuals), the latter two as the principal non-native contributors to admixed populations of the Americas. Genotype and allele frequency data of these populations are available in the online SNPforID browser <http://spsmart.cesga.es/snpforid.php?dataSet=snpforid52> [6]. Additionally, data from other admixed populations: Colombia-Medellín (60 individuals), Puerto Rico (55) and Mexico (66) was downloaded from the ENGINES genetic variant browser (<http://spsmart.cesga.es/engines.php?dataSet=engines>) [7]. Lastly, data from Colombian populations from Antioquia, Caldas, Eje Cafetero, Quindío, Risalda, Tolima, and Valle (totaling 210 individuals) was included using data from a previous study [8].

1.1. Ethical requirements

Informed written consent was obtained from all subjects. Ethical approval was granted by the institutes participating in Venezuela and Spain. This study followed the guidelines for publication of forensic population data [9].

1.2. Sampling, DNA extraction and SNP genotyping

Blood samples were collected on FTA cards and DNA was extracted by standard phenol/chloroform and salting out methods [11]. The 52-plex SNP typing assay followed the protocol previously described by Sanchez et al. [12], but using half PCR and extension reaction volumes in order to optimize the assay.

1.3. Quality control

After genotyping and automated analysis with GeneMapper™, data was then reviewed manually and results compared. This study has followed the ISFG recommendations for the use of SNP markers in the analysis of forensic population data, signifying the use of recommended nomenclature and guidelines regarding quality assurance and statistical issues [10].

1.4. Statistical analysis

Allele frequencies, observed and expected heterozygosities, Hardy–Weinberg (HW) equilibrium analysis and population pairwise comparisons were calculated with Arlequin (v. 3.5.1.2) [13]. Bonferroni multiple test corrections were applied to adjust the significance level of the HW tests. Estimation of forensic

informativeness metrics: random match probability and discrimination power plus paternity exclusion power and typical paternity indices were made using Promega PowerStats (<http://www.promega.com/geneticidtools/powerstats/>). *Structure* software (v.2.3.3) [14] was used to analyze population characteristics applying admixture and correlated frequencies models. Runs used 200,000 Markov Chain Monte Carlo steps after a burn-in length of 200,000. Five independent replicates were performed for each population cluster value (K : 2–5). A matrix corresponding to the average of the permuted matrices across replicates was estimated using CLUMPP [15], optimum K estimation used the online tool *Structure Harvester* [16] and visualization of cluster bar plots used *distruct* software [17].

2. Results and discussion

Data corresponding to allele frequencies, HW analyses, observed and expected heterozygosities, significant exact probabilities and standard deviations for the populations of Bari, Pemon, Panare and Warao are presented in [supplementary Table S2](#). Forensic informativeness metrics of matching probability, exclusion power and typical paternity indices are presented in [supplementary Table S3](#). After Bonferroni correction, p values <0.00019 were considered statistically significant, however no populations showed departures from HW equilibrium for any SNPs. Levels of heterozygosity were noticeably lower in the study populations and this is reflected in the forensic metrics with lower values when compared to Europeans, Africans and admixed populations from South America, as shown in [Table 1](#).

Pairwise comparisons between Europeans and between Africans with the study populations gave significant levels of divergence in all cases ($p < 0.01$). Among admixed populations, Ecuadorians showed the lowest distance to the study populations and Puerto Ricans the highest ([Table 2](#)).

As previously reported by Sanchez et al. [12], using the same 52 SNPs for cluster analysis of global population panels with *Structure*, revealed four genetic clusters corresponding to African, East Asian, European and Native American (Greenland) groups. We performed a similar analysis to evaluate the population clustering and the ancestry proportions of the groups considered in this work. However, it is important to note that this 52-plex assay was designed by selecting markers according to their heterozygosity for individual differentiation, therefore, the clustering observed between these populations is certain to be improved by the use of ancestral informative markers (AIMs).

The cluster plots from the *Structure* analysis performed on three equivalent populations (excluding East Asians) are summarized in [Fig. 1](#), indicating an optimum K value of three. Admixed

Table 1

Cumulative random match probability and exclusion probability of the four study populations compared with previously established data analyzed with the 52-plex system.

	Cumulative random match probability	Cumulative exclusion probability	Source
Europe	3.06×10^{-21}	0.999	SNPforID browser
Africa	5.96×10^{-18}	0.998	SNPforID browser
Admixed Ecuador ^a	3.30×10^{-17}	0.998	SNPforID browser [18]
Admixed Colombia I	2.97×10^{-20}	0.999	SNPforID browser [8]
Admixed Colombia II ^b	2.94×10^{-20}	0.999	ENGINES (SPSmart)
Admixed Mexico ^b	2.80×10^{-19}	0.999	ENGINES (SPSmart)
Admixed Puerto Rico ^b	4.01×10^{-20}	0.999	ENGINES (SPSmart)
Bari	1.88×10^{-13}	0.988	Present study
Pemon	3.83×10^{-15}	0.993	Present study
Panare	1.36×10^{-12}	0.982	Present study
Warao	1.86×10^{-15}	0.992	Present study
All study populations	2.52×10^{-15}	0.994	Present study

^a 49/52 SNPs analyzed.

^b 50/52 SNPs analyzed.

Please cite this article in press as: Y. Ruiz, et al., Analysis of the SNPforID 52-plex markers in four Native American populations from Venezuela, Forensic Sci. Int. Genet. (2012), doi:10.1016/j.fsigen.2012.02.007

Table 2

F_{st} values for pairwise comparisons of study and admixed populations of South America plus Europeans and Africans. All observed values were significant ($p < 0.01$).

	Bari	Panare	Pemon	Warao	All
European	0.17208	0.17536	0.12723	0.11494	0.13245
African	0.21426	0.22859	0.20948	0.18440	0.19037
Admixed Ecuador ^a	0.10585	0.10132	0.05056	0.05016	0.04690
Admixed Colombia I	0.13384	0.14393	0.09998	0.07392	0.09355
Admixed Colombia II ^b	0.13384	0.14393	0.09998	0.07392	0.09355
Admixed Puerto Rico ^b	0.19206	0.19549	0.15432	0.13100	0.15790
Admixed Mexico ^b	0.18070	0.17171	0.13114	0.11314	0.13512

^a 49/52 SNPs analyzed.

^b 50/52 SNPs analyzed.

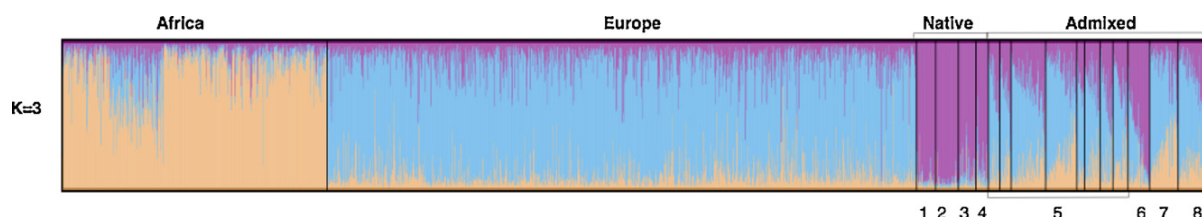


Fig. 1. Cluster analysis using *Structure* software for three assumed populations ($K = 3$): African (orange), European (blue) and Native American (purple). Study populations from Venezuela are represented as: Bari (1), Panare (2), Pemon (3), Warao (4). Admixed populations are represented as: Colombia (5), Ecuador (6), Puerto Rico (7), Mexico (8). Estimation of the optimum K value for this analysis is outlined in [supplementary Fig. S4](#). Admixed populations 5–8 are arranged in increasing second admixture component (purple Native American for 5, 6, 8 and orange African for 7).

populations showed a range of membership proportions in their ancestral contributions: Ecuadorians have a predominantly Amerindian component which confirm previous studies [18], this component is present at a reduced proportion in Mexicans, while Colombians showed a predominantly European component. Puerto Ricans showed the highest African component, in agreement with a previous study [19]. Note that among the African group used for comparison purposes, the Somalis (104 individuals) showed a significant admixed component – visible as a distinct set of columns in the middle of the African cluster of Fig. 1.

Our results support the suggested demographic history of the Native American populations we analyzed. These four Amerindian populations have existed as largely isolated groups as a consequence of their cultural patterns, geographic distances and linguistic differences [20]. A similar lack of European and African genetic contribution in the study populations, particularly in Warao, Pemon and Panare, has been shown in a recent study using a larger set of 450 ancestry informative marker SNPs, which included several other Native South American populations [21]. A previous genetic study using STR markers also revealed a clear genetic differentiation of the Bari population compared to other neighboring populations [20].

The implementation of population data, as presented in this work, is of interest for application in forensic analysis and paternity cases, in which allele frequency estimation for local databases is required. South American populations are of particular interest because their demographic structures have changed dramatically since the original peopling of the continent, and as a consequence, today they show a wide range of admixture patterns from European, African and Native American genetic components. It is worth recording that many geographically isolated populations in South America, such as those of this study, have managed to survive in a state of genetic preservation and therefore represent an important resource for the analysis of unadmixed Native American population variability.

Acknowledgments

The authors thank all the Venezuelan institutions involved in the populations sampling, and for technical support provided by Ana Freire and Carla Santos at the Forensic Genetics Unit,

University of Santiago de Compostela. YR was supported by the Fundation *Gran Mariscal de Ayacucho* (FUNDAYACUCHO). MVL was supported by funding from Xunta de Galicia INCITE 09 208163PR and this work was in part supported by additional funding from Xunta de Galicia: PGIDTIT06P-XIB228195PR. AC was supported by FIS PS09/02368 (FEDER funding).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.fsigen.2012.02.007](https://doi.org/10.1016/j.fsigen.2012.02.007).

References

- [1] A.L. Bryan, R.M. Casamiquela, J.M. Cruxent, R. Gruhn, C. Ochsenius, An El Jobo Mastodon kill at Taimatima, Venezuela, *Science* 200 (1978) 1275–1277.
- [2] G. Morón, Historia de Venezuela, Italgáfica, 1971.
- [3] N. Pocater, D. Egildo-Palau, Proyecto ley de patrimonio cultural de los pueblos y comunidades indígenas, Asamblea Nacional, Artículo 1, 2007.
- [4] C. Phillips, M. Fondevila, M. Garcia-Magarinos, A. Rodriguez, A. Salas, A. Carracedo, M.V. Lareu, Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers, *Forensic Sci. Int. Genet.* 2 (2008) 198–204.
- [5] M. Fondevila, C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, A. Carracedo, M.V. Lareu, Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, *Forensic Sci. Int. Genet.* 2 (2008) 212–218.
- [6] J. Amigo, C. Phillips, M.V. Lareu, A. Carracedo, The SNPforID browser: an online tool for query and display of frequency data from the SNPforID project, *Int. J. Legal Med.* 122 (2008) 435–440.
- [7] J. Amigo, A. Salas, C. Phillips, ENIGES: exploring single nucleotide variation in entire human genomes, *BMC Bioinformatics* 12 (2011) 105.
- [8] L. Porras, C. Phillips, M. Fondevila, L. Beltran, T. Ortiz, F. Rondon, G. Barreto, M.V. Lareu, J. Henao, A. Carracedo, Genetic variability of the SNPforID 52-plex identification-SNP panel in Central West Colombia, *Forensic Sci. Int. Genet.* 4 (2009) e9–e10.
- [9] A. Carracedo, J.M. Butler, L. Gusmao, W. Parson, L. Roewer, P.M. Schneider, Publication of population data for forensic purposes, *Forensic Sci. Int. Genet.* 4 (2010) 145–147.
- [10] P.M. Schneider, Scientific standards for studies in forensic genetics, *Forensic Sci. Int.* 165 (2007) 238–243.
- [11] J. Sambrook, D.W. Russell, *Molecular Cloning: A Laboratory Manual*, Harbor Laboratory Press, 2001.
- [12] J.J. Sanchez, C. Phillips, C. Borsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P.M. Schneider, A. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, *Electrophoresis* 27 (2006) 1713–1724.

- [13] L. Excoffier, G. Laval, S. Schneider, Arlequin (version 3.0): an integrated software package for population genetics data analysis, *Evol. Bioinform. Online* 1 (2005) 47–50.
- [14] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, *Genetics* 155 (2000) 945–959.
- [15] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, *Bioinformatics* 23 (2007) 1801–1806.
- [16] D.A. Earl, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method, *Conservation Genetics Resources*.
- [17] N.A. Rosenberg, DISTRUCT: a program for the graphical display of population structure, *Mol. Ecol. Notes* 4 (2004) 137–138.
- [18] L. Poulsen, C. Borsting, C. Tomas, F. Gonzalez-Andrade, R. Lopez-Pulles, J. Gonzalez-Solorzano, N. Morling, Typing of Amerindian Kichwas and Mestizos from Ecuador with the SNPforID multiplex, *Forensic Sci. Int. Genet.* 5 (2011) e105–e107.
- [19] K. Bryc, C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C.D. Bustamante, H. Ostrer, Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations, *Proc. Natl. Acad. Sci. U. S. A.* 107 (Suppl. 2) (2010) 8954–8961.
- [20] W.M. Zabala Fernandez, L. Borjas-Fajardo, E. Fernandez Salgado, C. Castillo, L. Socca, M.G. Portillo, M.A. Sanchez, W. Delgado, A. Morales-Machin, Z. Layrisse, L. Pineda Bernal, Use of short tandem repeats loci to study the genetic structure of several populations from Zulia State, Venezuela, *Am. J. Hum. Biol.* 17 (2005) 451–459.
- [21] J.M. Galanter, J.C. Fernandez, C. Gignoux, J. Barnholtz-Sloan, C. Fernandez, M. Via, A. Hidalgo-Miranda, A.V. Contreras, L.U. Figueroa, P. Raska, G. Jimenez-Sanchez, I.S. Zolezzi, M. Torres, C.R. Ponte, Y. Ruiz, A. Salas, E. Nguyen, C. Eng, L. Borjas, W. Zabala, G. Barreto, F. Rondón, A. Ibarra, P. Taboada, L. Porras, F. Moreno, A. Bingham, G. Guitierrez, T. Brutsearet, F.L. Velarde, L. Moore, E. Vargas, M. Cruz, J. Escobedo, J. Rodriguez-Cintrón, R. Chapela, J.G. Ford, C. Bustamante, D. Seminaria, M. Shriver, E. Ziv, E.G. Burchard, E. Parra, A. Carracedo, Development of a Panel of Genome-wide Ancestry Informative Markers to Study Admixture Throughout the Americas, *PloS Genetics* 8 (2012), e1002554.

Bloque 1.

V.2. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas

Joshua Mark Galanter, Juan Carlos Fernandez, Christopher R. Gignoux, Jill Barnholtz Sloan, Ceres Fernandez, Marc Via, Alfredo Hidalgo-Miranda, Alejandra V. Contreras, Laura Uribe Figueroa, Paola Raska, Gerardo Jimenez-Sanchez, Irma Silva Zolezzi, Maria Torres, Clara Ruiz Ponte, Yarimar Ruiz, Antonio Salas, Elizabeth Nguyen, Celeste Eng, Lisbeth Borjas, William Zabala, Guillermo Barreto, Fernando Rondón, Adriana Ibarra, Patricia Taboada, Liliana Porras, Fabián Moreno, Abbigail Bigham, Gerardo Gutierrez, Tom Brutsaert, Fabiola León-Velarde, Lorna G. Moore, Enrique Vargas, Miguel Cruz, Jorge Escobedo, Jose Rodriguez-Santana, William Rodriguez-Cintrón, Rocio Chapela, Jean G. Ford, Carlos Bustamante, Daniela Seminara, Mark Shriver, Elad Ziv, Esteban Gonzalez Burchard, Robert Haile, Esteban Parra, Angel Carracedo for the LACE consortium.

(*PloS Genetics*, 2012, 8(3): e1002554. doi:10.1371/journal.pgen.1002554)

Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas

Joshua Mark Galanter^{1*}, Juan Carlos Fernandez-Lopez², Christopher R. Gignoux¹, Jill Barnholtz-Sloan³, Ceres Fernandez-Rozadilla⁴, Marc Via⁵, Alfredo Hidalgo-Miranda², Alejandra V. Contreras², Laura Uribe Figueroa², Paola Raska³, Gerardo Jimenez-Sanchez², Irma Silva Zolezzi², Maria Torres⁴, Clara Ruiz Ponte⁴, Yarimar Ruiz⁴, Antonio Salas⁴, Elizabeth Nguyen¹, Celeste Eng¹, Lisbeth Borjas⁶, William Zabala^{4,6}, Guillermo Barreto⁷, Fernando Rondón González⁸, Adriana Ibarra⁹, Patricia Taboada^{4,10}, Liliana Porras^{4,11}, Fabián Moreno¹², Abigail Bigham¹³, Gerardo Gutierrez¹⁴, Tom Brutsaert¹⁵, Fabiola León-Velarde¹⁶, Lorna G. Moore¹⁷, Enrique Vargas¹⁸, Miguel Cruz¹⁹, Jorge Escobedo²⁰, José Rodríguez-Santana²¹, William Rodríguez-Cintrón²², Rocio Chapela²³, Jean G. Ford²⁴, Carlos Bustamante²⁵, Daniela Seminara²⁶, Mark Shriver²⁷, Elad Ziv¹, Esteban Gonzalez Burchard¹, Robert Haile²⁸, Esteban Parra²⁹, Angel Carracedo^{4,9}, for the LACE Consortium

1 University of California San Francisco, San Francisco, California, United States of America, **2** Instituto Nacional de Medicina Genómica, Mexico City, Mexico, **3** Case Western Reserve University, Cleveland, Ohio, United States of America, **4** Fundación Pública Galega de Medicina Xenómica (SERGAS)-CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain, **5** Universitat de Barcelona, Barcelona, Spain, **6** Universidad del Zulia, Maracaibo, Venezuela, **7** Universidad del Valle, Santiago de Cali, Colombia, **8** Universidad Industrial de Santander, Bucaramanga, Colombia, **9** Universidad de Antioquia, Medellín, Colombia, **10** Instituto de Investigaciones Forenses, Sucre, Bolivia, **11** Universidad Tecnológica de Pereira, Pereira, Colombia, **12** Unidad de Genética Forense, Servicio Médico-Legal de Chile, Santiago de Chile, Chile, **13** University of Michigan, Ann Arbor, Michigan, United States of America, **14** University of Colorado at Boulder, Boulder, Colorado, United States of America, **15** Syracuse University, Syracuse, New York, United States of America, **16** Universidad Peruana Cayetano Heredia, Lima, Peru, **17** Wake Forest University, Winston-Salem, North Carolina, United States of America, **18** Universidad Mayor de San Andrés, La Paz, Bolivia, **19** Centro Médico Nacional Siglo XXI, Mexican Social Security Institute (IMSS), Mexico City, Mexico, **20** Hospital General Regional 1, IMSS, Mexico City, Mexico, **21** Centro de Neumología Pediátrica, San Juan, Puerto Rico, **22** VA Caribbean Health System, San Juan, Puerto Rico, **23** Instituto Nacional de Enfermedades Respiratorias (INER), Mexico City, Mexico, **24** Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America, **25** Stanford University, Stanford, California, United States of America, **26** National Cancer Institute, Bethesda, Maryland, United States of America, **27** Penn State University, University Park, Pennsylvania, United States of America, **28** University of Southern California, Los Angeles, California, United States of America, **29** University of Toronto at Mississauga, Mississauga, Canada

Abstract

Most individuals throughout the Americas are admixed descendants of Native American, European, and African ancestors. Complex historical factors have resulted in varying proportions of ancestral contributions between individuals within and among ethnic groups. We developed a panel of 446 ancestry informative markers (AIMs) optimized to estimate ancestral proportions in individuals and populations throughout Latin America. We used genome-wide data from 953 individuals from diverse African, European, and Native American populations to select AIMs optimized for each of the three main continental populations that form the basis of modern Latin American populations. We selected markers on the basis of locus-specific branch length to be informative, well distributed throughout the genome, capable of being genotyped on widely available commercial platforms, and applicable throughout the Americas by minimizing within-continent heterogeneity. We then validated the panel in samples from four admixed populations by comparing ancestry estimates based on the AIMs panel to estimates based on genome-wide association study (GWAS) data. The panel provided balanced discriminatory power among the three ancestral populations and accurate estimates of individual ancestry proportions ($R^2 > 0.9$ for ancestral components with significant population-between-subject variance). Finally, we genotyped samples from 18 populations from Latin America using the AIMs panel and estimated variability in ancestry within and between these populations. This panel and its reference genotype information will be useful resources to explore population history of admixture in Latin America and to correct for the potential effects of population stratification in admixed samples in the region.

Citation: Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, et al. (2012) Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas. *PLoS Genet* 8(3): e1002554. doi:10.1371/journal.pgen.1002554

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received: September 24, 2011; **Accepted:** January 10, 2012; **Published:** March 8, 2012

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: This work was made possible by the Latin American Cancer Epidemiology (LACE) Consortium, which has been supported with travel/meeting grants from the National Cancer Institute and by Contract #HHSN261201000641P from the National Cancer Institute, National Institutes of Health. The GALA study was funded by the NIH (R01HL078885, K23HL004464, 5R01HL088133), the American Asthma Foundation, and the Sandler Foundation. The Mexico City type 2 diabetes study was funded in Canada by the Canadian Institutes of Health Research, the Banting and Best Diabetes Centre, the Canada Foundation for Innovation, and The Ontario Innovation Trust, and in Mexico by the following grants: CONACYT SALUD-2005-C02-14412 and 2007-C01-71068, Proyectos Estratégicos, Apoyo Financiero Fundación IMSS, and Fundación Gonzalo Río Arronte I, A.P. Mexico. Samples in Latin America were collected and typed with funding support from FIS P509/02368 (FEDER funding). Samples in Bolivia were collected with funding support from the National Institutes of Health (R01HL079647). JMG received support from the National Institutes of Health (T32GM007546, 2KL2RR024130) and the Ralph Hewitt Fellowship. EP is the recipient of a CIHR New Investigator Award. CRG was supported in part by NIH Training Grant T32GM007175. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: CB and EGB are on the Scientific Advisory Board of 23andme's project "Roots into the Future." 23andme is a Mountain View, CA, company that provides direct-to-consumer genetic products. CB is also on the SAB of Ancestry.com, a company in Provo, UT, that provides direct-to-consumer genetic products. Neither company played a part in the research described here or has a financial stake in the results. The remaining authors have declared that no competing interests exist.

* E-mail: joshua.galanter@ucsf.edu

† These authors contributed equally to this work.

Introduction

Most individuals from the Americas are admixed descendants of Native American, European, and African ancestors. Complex historical factors have resulted in varying proportions of ancestral contributions between individuals within and between ethnic groups [1]. For example, in a study of five Hispanic/Latino ethnic groups, Puerto Ricans and Dominicans showed the largest proportion of African ancestry, while Mexicans had a significantly larger proportion of Native American ancestry than the other groups [2]. Even within small islands in the Caribbean there can be high variance in admixture proportions [3]. Ancestry Informative Markers (AIMs) are commonly used to estimate overall admixture proportions efficiently and inexpensively [4]. AIMs are polymorphisms that exhibit large allele frequency differences between populations and can be used to infer individuals' geographic origins. For example, the forensic use of a panel of AIMs successfully identified the ancestral origin of seven unmatched samples implicated in the 11-M Madrid commuter train bombings of 2004 [5]. Using a panel of AIMs distributed throughout the genome, it is possible to estimate the relative ancestral proportions in admixed individuals such as African Americans and Latin Americans, as well as to infer the time since the admixture process [6,7].

In addition to providing estimates of individual's ancestral history, admixture proportions can be correlated to physiologic measurements such as spirometric measurements of lung function [8] and uterine artery blood flow [9], risk of diseases such as peripheral vascular disease [10] and breast cancer [11], as well as to control for the effects of population stratification in genetic association studies [12]. Consequently, it is important for researchers to have access to validated, accurate panels of AIMs that can be used for Latin American populations throughout the Americas, including Hispanics/Latinos in the United States, where according to the US census bureau, they are the fastest growing ethnic group [13].

Several groups have described panels of AIMs designed to estimate individual ancestry and to control for the effects of population stratification in Latino populations [14,15,16,17]. However, in most cases these studies were limited in the number of AIMs selected, lack of systematic basis for the selection of AIMs, and lack of validation compared to robust estimates of ancestry based on genome-wide data from hundreds of thousands of markers. Additionally, most published AIMs panels lack availability of genotyping data of relevant ancestral populations.

In this paper, we describe a three-stage approach to developing a panel of 446 Ancestry Informative Markers (AIMs) optimized to characterize admixture throughout Latin America. In the first stage,

we used genome-wide data from two African populations, three European populations, and six Native American populations to select AIMs that were informative, evenly distributed throughout the genome, and portable, having little within-continent heterogeneity. In the second stage, we validated the panel of AIMs in four admixed samples by comparing the ancestry estimates based on the AIMs panel with ancestry estimates based on genome-wide data. In the final stage, using these AIMs, we genotyped samples from 18 additional populations originating throughout the Americas to estimate ancestry differences within and between populations and to determine the onset of admixture for each group.

Results

AIMs selection

A total of 446 AIMs were identified; the panel is presented in its entirety in Table S1. The 400 most informative markers were used to design multiplexes for the Sequenom genotyping platform. Consistent with the goals of the study, the AIMs panel provides a balanced set of markers capable of distinguishing the three ancestral populations of modern Latin Americans. Specifically, the cumulative locus-specific branch length for the I_n statistic was 43.8, 44.0, and 44.0 for Africans, Europeans, and Native Americans, respectively. Because the mean locus specific branch length for European ancestry was lower than for African or Native American ancestry, there are 202 European AIMs with a median LSBL F_{st} of 0.37 (25: 75 percentiles 0.35–0.41) and a median LSBL I_n of 0.21 (25:75 percentiles 0.20–0.23). There are 115 African AIMs with a median LSBL F_{st} of 0.63 (25: 75 percentiles 0.61–0.66) and a median LSBL I_n of 0.37 (25:75 percentiles 0.36–0.40). The 129 Native American AIMs have a median LSBL F_{st} of 0.56 (25: 75 percentiles 0.54–0.61) and a median LSBL I_n of 0.33 (25:75 percentiles 0.32–0.36). The informativeness of the AIMs panel is summarized in Table 1. The lower informativeness of European-specific AIMs is likely because European populations are geographically and genetically intermediate to African and Native American populations [18,19]. Consequently, more European AIMs were needed to provide balanced discriminatory power.

Validation of the panel of AIMs

Figure 1 shows the accuracy of the ancestry estimates obtained with the AIMs panel for four Latin American samples. The individual ancestry estimates based on the AIMs panel were compared to the estimates based on genome-wide data. Generally, there was strong concordance between ancestry estimates using the AIMs and using GWAS data. There is a slight systematic underestimate of European ancestry in all four populations tested, and a slight overestimate of African ancestry. Table 2 summarizes

Author Summary

Individuals from Latin America are descendants of multiple ancestral populations, primarily Native American, European, and African ancestors. The relative proportions of these ancestries can be estimated using genetic markers, known as ancestry informative markers (AIMs), whose allele frequency varies between the ancestral groups. Once determined, these ancestral proportions can be correlated with normal phenotypes, can be associated with disease, can be used to control for confounding due to population stratification, or can inform on the history of admixture in a population. In this study, we identified a panel of AIMs relevant to Latin American populations, validated the panel by comparing estimates of ancestry using the panel to ancestry determined from genome-wide data, and tested the panel in a diverse set of populations from the Americas. The panel of AIMs produces ancestry estimates that are highly accurate and appropriately controlled for population stratification, and it was used to genotype 18 populations from throughout Latin America. We have made the panel of AIMs available to any researcher interested in estimating ancestral proportions for populations from the Americas.

the performance of ancestry estimates for in the four admixed samples. The correlation (R^2) between ancestry estimates using AIMs and ancestry from GWAS data is high in most cases, especially for Native American and European ancestry in all three Mexican samples and European and African ancestry in the Puerto Rican sample. The correlation coefficient was lower for estimates of ancestry where there was less variance in the true ancestral proportion.

Use of AIMs panel subsets to control for population stratification

We investigated the effect of the number of AIMs on the accuracy of the estimates of ancestry, using the parents of Puerto Rican subjects with asthma from the GALA study ($n = 803$) [20] and the sample from Mexico City, which includes 967 cases and 343 controls from a Type II Diabetes study [21,22]. We compared the estimates of ancestry based on genome-wide data with the estimates obtained with different subsets of AIMs (314, 194, 88, 41 and 22). For this analysis, we first started with the 314 AIMs that were genotyped in this sample. We produced nested subsets of AIMs by progressively reducing the number of AIMs, keeping only the most informative markers, and ensuring that the final panel of AIMs was balanced (e.g. each panel has approximately the same

ancestry information content for each ancestral group). Ancestry estimates were estimated with the program ADMIXTURE with ancestral genotype data. Table 3 and Figure 2 depict the correlation (R^2) between the genome-wide estimates and the estimates based on the panel of AIMs, as well as the mean differences, mean absolute differences and root mean square errors. As expected, reducing the number of AIMs in the panel results in decreasing correlation and increasing error of the ancestry estimates compared to the estimates produced with genome-wide data. Performance of the 194 AIMs panel, and to a lesser extent the 88 AIMs panel is comparable to performance of the 314 AIMs panel. The correlations between the estimates based on 22 AIMs and those based on genome-wide estimates are considerably worse, particularly for the estimates of African ancestry in Mexicans and Native American ancestry in Puerto Ricans, which are the ancestral components with the least amount of variance between subjects.

We evaluated the utility of the different panels of AIMs to control for the effects of population stratification in the Mexico City sample, which had previously been shown to have significant population stratification [22]. The average Native American ancestry in the cases was estimated to be 66% versus 57% in the control group. We carried out a logistic regression analysis to test the association of approximately 315,000 common markers with type 2 diabetes, including as covariates sex and age, or alternatively, sex, age and the ancestry estimates obtained with 314, 194, 88, 41 and 22 AIMs. We then prepared quantile-quantile (QQ) plots comparing the p values obtained in the logistic association tests with the values expected under the null model of no association (See Figure S1). The extent of population stratification was quantified by the inflation factor lambda [23], using the program WGAViewer. Under the model conditioning by sex and age, there was a strong departure of the observed and expected p-values. The value of lambda was 1.4, indicating grossly inflated false-positive rates. As seen in Figure 2, adding ancestry estimates to the model dramatically reduced the inflation factor: reducing lambda to 1.04 using genome-wide estimates of ancestry. Using AIMs panels of 314 AIMs, 194 AIMs and 88 AIMs produced nearly equal reductions in lambda. Performance using smaller AIMs panels still resulted in a marked decrease in the inflation factor: for the 41 and 21 AIMs panels lambda was 1.05.

Ancestry estimates for 18 populations in the Americas

The panel of AIMs was carried forward to genotype a total of 373 individuals from 18 populations throughout the Americas using the Sequenom platform. Generally speaking, the platform performed well, though 75 SNPs were excluded due to lower call rates (all samples included). The final analysis was based on 325 markers. Among the SNPs meeting quality control criteria,

Table 1. Characteristics of the AIMs panel.

Population	Number of AIMs	Cumulative LSBL F_{st}	Cumulative LSBL I_n	LSBL F_{st}	LSBL I_n
				(mean \pm sd; median, 25:75)	(mean \pm sd; median, 25:75)
African	115	73.0	43.8	0.64 \pm 0.05; 0.63, 0.61: 0.66	0.38 \pm 0.03; 0.37, 0.36: 0.40
European	202	77.9	44.0	0.39 \pm 0.05; 0.37, 0.35: 0.41	0.22 \pm 0.03; 0.21, 0.20: 0.23
Native American	129	74.5	44.0	0.58 \pm 0.05; 0.56, 0.54: 0.61	0.34 \pm 0.03; 0.33, 0.32: 0.36

doi:10.1371/journal.pgen.1002554.t001

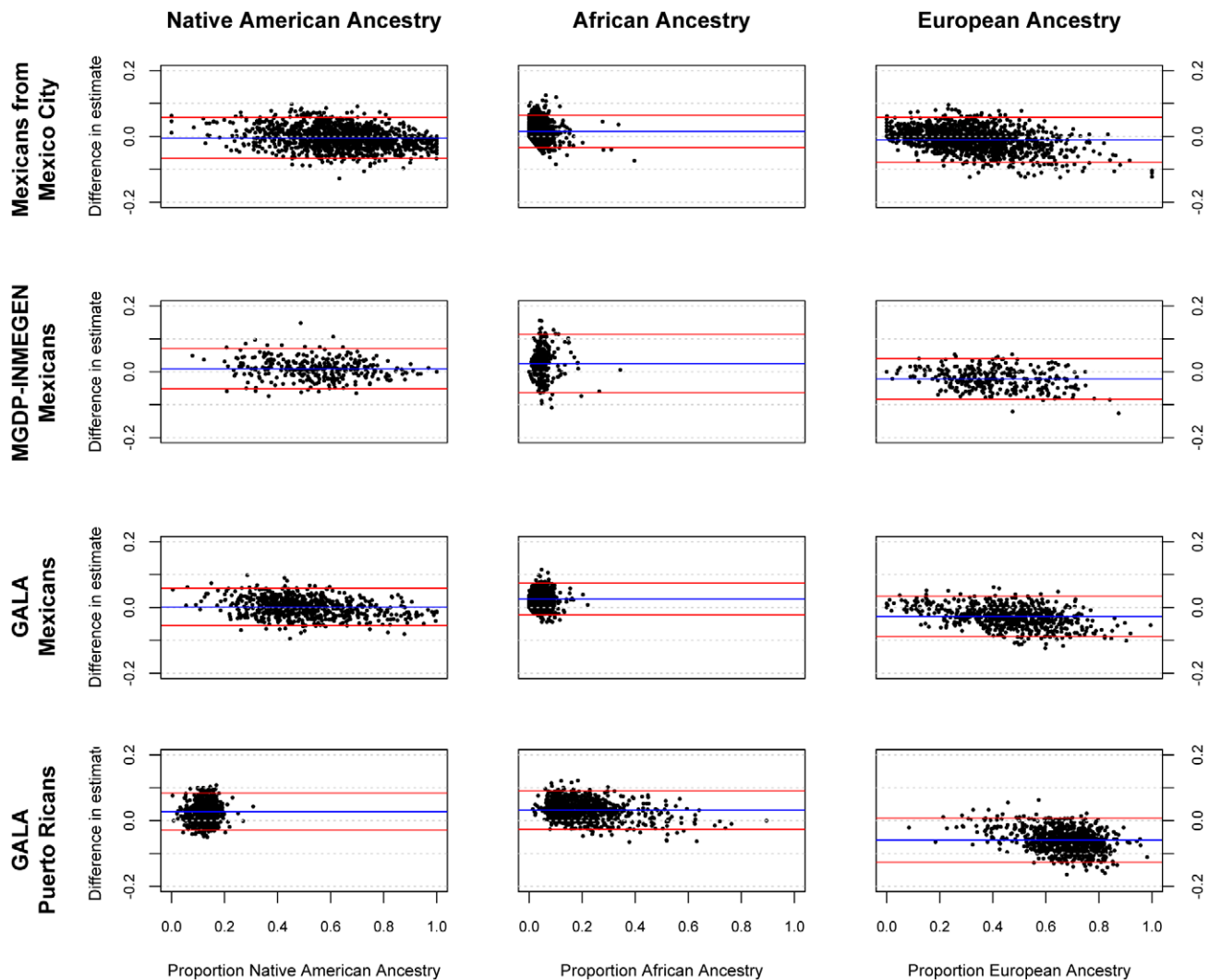


Figure 1. Bland-Altman plots showing error in individual ancestral estimates using AIMs to ancestral estimates using GWAS data. The x-axis shows the ancestry estimate using GWAS data; the y-axis shows the difference in estimates between GWAS and AIMs data using the 425 AIMs genotyped in the GALA Mexicans and Puerto Rican samples, 314 AIMs for the Mexico City sample, and 398 AIMs for the MGDP-INMEGEN sample.

doi:10.1371/journal.pgen.1002554.g001

the average call rate was 91.7% (max value 99.5% and min value 55.1%, all samples included). Two additional populations (Coyas and Mapuches) were genotyped but excluded from the analysis due to the low quality of the samples. Four additional individuals were excluded due to genotyping call rates of <90%.

Table 4 summarizes the ancestral estimates obtained for the 18 populations characterized and Figure 3 shows one-dimensional scatter plots of ancestry for each ancestral component in each sample. As expected, most of the indigenous populations have high Native American ancestry, with a median (25:75 percentile) Native American ancestry of 0.80 (0.57: 0.87) for Colombian Awa, 0.86 (0.83: 0.89) for Colombian Coyaima, 0.83 (0.64: 0.87) for Colombian Pastos, 1.0 (1.0: 1.0) for Venezuelan Panare and Pemón, 0.99 (0.97: 1.0) for Venezuelan Warao, and 0.97 (0.84: 1.0) for Venezuelan Wayu. The Wichi from Argentina had relatively lower Native American ancestry, which was estimated as 0.41 (0.12: 0.84). The Bolivian individuals recruited in the Beni and Cochabamba Departments, as well as those from the

Altiplano region of the La Paz Department also showed high Native American ancestry proportions. The median Native American ancestries for these samples were 0.94 (0.78: 0.96), 0.90 (0.86: 0.95) and 0.98 (0.96: 1.0), respectively. In contrast, in the Bolivian sample from the subtropical Yungas region, which is known for the presence of scattered Afro-Bolivian communities, many individuals had relatively high African ancestry (>0.6), whereas other individuals showed primarily Native American ancestry (>0.8) (Table 4 and Figure 3). The median African and Native American ancestries observed in the Yungas sample were 0.70 (0.01: 0.82) and 0.25 (0.13: 0.97), respectively.

The two Afro-Colombian samples included in this study had a median African ancestry of 0.76 (0.64: 0.83) for Chocó and 0.54 (0.46: 0.69) for Mulaló. Finally, the Mestizo samples from Colombia, Venezuela and Northern and Southern Chile showed a relatively high dispersion in Native American and European admixture proportions. With the exception of some Venezuelans from Maracaibo, on the Caribbean coast, most of these individuals had small (<10%) African contributions.

Table 2. Validation of the AIMs panel compared to ancestry estimates using GWAS data.

Sample Ancestry	Mean ancestry estimate (with GWAS)	Correlation R ²	Mean error \pm sd	Mean discordance	Root mean square error
Mexico City					
Native American	0.642	0.968	−0.005 (\pm 0.032)	0.025	0.032
European	0.324	0.956	−0.010 (\pm 0.034)	0.028	0.036
African	0.035	0.555	0.015 (\pm 0.025)	0.023	0.029
MGDP-INMEGEN					
Native American	0.544	0.966	0.009 (\pm 0.031)	0.025	0.032
European	0.402	0.964	−0.022 (\pm 0.031)	0.031	0.038
African	0.054	0.722	0.012 (\pm 0.023)	0.020	0.026
Mexico GALA					
Native American	0.496	0.972	0.002 (\pm 0.029)	0.023	0.029
European	0.458	0.967	−0.027 (\pm 0.031)	0.033	0.041
African	0.046	0.558	0.026 (\pm 0.025)	0.029	0.035
Puerto Rico GALA					
Native American	0.124	0.603	0.027 (\pm 0.029)	0.033	0.040
European	0.670	0.914	−0.059 (\pm 0.034)	0.060	0.068
African	0.206	0.942	0.032 (\pm 0.030)	0.036	0.044

doi:10.1371/journal.pgen.1002554.t002

We also estimated the average number of generations since admixture for the Mestizo and African descendant samples (Figure 4). Generally speaking, in the Afro-Colombian and Afro-Bolivian samples, the estimated time since admixture was 6.7 generations (95% credible interval: 5.4–8.4) for the Yungas, 5.8 generations (95% credible interval: 5.0–6.6) for the Mulaló and 7.34 generations (95% credible interval: 6.3–8.4) for the Chocó. In contrast, the estimates of time since admixture for the Mestizo samples were higher; the estimated time since admixture was 8.4 generations (95% credible interval: 6.9–10.3) for the Northern Chileans, 9.6 generations (95% credible interval: 7.9–11.8) for the Southern Chileans, 12.9 generations (95% credible interval: 10.5–16.0) for the Colombians, and 9.7 generations (95% credible interval: 7.9–12.1) for the Venezuelans.

Discussion

In this study, we developed, validated, and tested a novel panel of AIMs designed to accurately estimate the ancestral components (African, European, and Native American) of contemporary Latin American populations. We developed a new algorithm (provided in the web resources online) capable of taking genome-wide data from multiple populations within each continental group and identifying the most informative, well-balanced and portable markers to estimate ancestry proportions.

The ancestral samples used to identify the AIMs represented a wide variety of populations within each continental group. Specifically, we used six samples from Mesoamerica and the South American Andes as representatives of the ancestral Native American populations that make up modern Latin Americans. Our Native American samples had a median Native American ancestry of 97.7% (25: 75 range 93.2% to 100%) based on ancestry ascertainment using genomewide data. Given the history of European colonization in the Americas, a small amount of European genetic admixture (2.3%, 1×10^{-5} : 6.2%) is not surprising. However, a small amount of European admixture

would be expected to result in an underestimate of the information content of our AIMs. Although we did not include Native American populations from English-speaking North America for our analysis, our selection of markers excluded those with significant heterogeneity between Native American populations. Thus, we have no reason to believe the markers cannot be applied to North American populations, though the use of these markers for populations outside of Latin America should be pursued with caution.

We also included two samples from Africa (Yoruba from Nigeria and Luhya from Kenya, in East Africa). Historical records and genetic analyses indicate that most of the slaves imported into the Americas originated in West Africa [24]. Although it would have been ideal to include multiple West African ancestral populations, we included the Luhya sample in our study because unlike the Yoruba, who are descendants of the Benue-Congo subfamily of the Niger-Congo language family, the Luhya are a Bantu-speaking population, and many of the enslaved Africans brought to the Americas were Bantu speakers. Multiple studies show that the Luhya and other Bantu-speaking groups from East Africa are more closely related to West African Bantu speakers than to other East African ethnic groups [24,25]. In addition, a small but significant number of slaves originated in Southeastern Africa [26,27,28]. Finally, we used three European samples to estimate ancestral frequencies in Europe. Importantly, samples from Italy and the Iberian Peninsula, which have been the largest sources of European migrants to Latin America, were included in this analysis.

By excluding markers with significant within-continent heterogeneity, the selected panel of AIMs should be broadly portable to populations from throughout the Americas. Moreover, the exclusion of markers exhibiting substantial within-continent heterogeneity serves to ensure that there is relatively little bias in the estimates of ancestral allele frequency. This is because any bias would have had to occur in all of the ancestral populations within a given continent, at a similar magnitude and in the same direction. On the other hand, by design, the AIMs panel would not

Table 3. Performance of nested subsets of AIMs.

Sample	Correlation R ²	Mean error	Mean discordance	RMSE
314 AIMs in Mexico City Mexicans				
Native American	0.97	−0.005	0.025	0.032
European	0.96	−0.010	0.028	0.036
African	0.56	0.015	0.023	0.029
314 AIMs in GALA Puerto Ricans				
Native American	0.54	0.025	0.034	0.042
European	0.89	−0.061	0.063	0.072
African	0.92	0.035	0.041	0.049
194 AIMs in Mexico City Mexicans				
Native American	0.95	−0.005	0.031	0.039
European	0.94	−0.011	0.033	0.042
African	0.48	0.016	0.026	0.034
194 AIMs in GALA Puerto Ricans				
Native American	0.43	0.025	0.034	0.042
European	0.85	−0.060	0.063	0.072
African	0.89	0.035	0.044	0.053
88 AIMs in Mexico City Mexicans				
Native American	0.92	−0.006	0.040	0.051
European	0.89	−0.014	0.044	0.056
African	0.35	0.020	0.034	0.044
88 AIMs in GALA Puerto Ricans				
Native American	0.27	0.035	0.052	0.064
European	0.72	−0.067	0.077	0.093
African	0.77	0.032	0.055	0.069
41 AIMs in Mexico City Mexicans				
Native American	0.85	−0.011	0.056	0.070
European	0.80	−0.016	0.061	0.076
African	0.21	0.027	0.044	0.059
41 AIMs in GALA Puerto Ricans				
Native American	0.14	0.038	0.069	0.086
European	0.56	−0.086	0.101	0.123
African	0.64	0.049	0.076	0.096
22 AIMs in Mexico City Mexicans				
Native American	0.76	−0.011	0.075	0.094
European	0.69	−0.027	0.081	0.103
African	0.14	0.038	0.059	0.081
22 AIMs in GALA Puerto Ricans				
Native American	0.10	0.041	0.086	0.108
European	0.39	−0.099	0.125	0.156
African	0.48	0.057	0.101	0.127

doi:10.1371/journal.pgen.1002554.t003

be expected to differentiate within-continent population substructure. Indeed, we found that the eight Native American populations genotyped with the AIMs panel were indistinguishable in principal component space beyond the first principal component, which represented the degree of European admixture (data not shown). There are several reasons we chose to exclude markers that could have potentially been used to differentiate within-continent substructure. First, the principal reason for designing this panel

was for identifying continental ancestry proportions in admixed samples, as continental admixture is the most important source of population structure in Latin Americans. Secondly, because we had a limited number of Native American ancestral groups available for study, we would have only been able to generate AIMs that distinguished Mesoamerican populations from Andean populations. Third, the use of heterogeneity filters was an important element of quality control, as it served to filter out alleles with extreme frequencies due to bias. Finally, because the genetic differences within continental groups are smaller than between continental groups, we would have required many more markers to accurately determine within-continent substructure.

We validated the panel of AIMs by comparing ancestry estimates derived from the subset of AIMs to estimates derived from genome-wide data in four Latin American populations, three from Mexico and one from Puerto Rico. Overall, the ancestral estimates for both Puerto Ricans and all Mexican groups were consistent with previously published literature [29]. Specifically, Bryc et al found that Puerto Ricans had $23.6\% \pm 12\%$ African ancestry, consistent with our finding of $20.6\% \pm 12.3\%$ and Mexicans had $5.6\% \pm 2\%$ African ancestry, consistent with our findings of between $3.5\% \pm 3.1\%$ and $5.4\% \pm 3.6\%$. The Native American component in the three Mexican populations ($64.2\% \pm 17.6\%$, $54.4\% \pm 16.9\%$, and $49.6\% \pm 17.4\%$ in Mexicans from Mexico City, INMEGEN, and GALA studies, respectively) is also consistent with results obtained by Bryc et al ($50.1\% \pm 13\%$) and in a study by Silva-Zolezzi et al of diverse Mexican Mestizo populations ($55.2\% \pm 15.4\%$) [30].

There was strong correlation between ancestral estimates obtained from the AIMs panel and those obtained from GWAS, providing strong support for the use of the AIMs panel to accurately estimate ancestry. For over 95 percent of the samples, the estimates of ancestry using AIMs were within 10% of the value obtained using GWAS data.

The correlation was lower for the minor ancestral components (African ancestry in Mexican populations and Native American ancestry in Puerto Rican populations). This reflected the more limited between-subject variance in the minor ancestral component. Since the coefficient of determination (R^2) represents the proportion of variance in the outcome variable (in this case, the true measure of ancestry), explained by the predictor (estimates of ancestry using AIMs), in cases where there is more limited variance in the outcome variable such as estimates of African ancestry in Mexicans and Native American ancestry in Puerto Ricans, we observe a lower R^2 . Nonetheless, measures of individual error in estimate, such as the root mean squared error, are comparable for all three ancestral estimates in both Puerto Ricans and Mexicans, suggesting that the panel performs consistently across all ancestral components, and in most cases, the estimate of ancestry using AIMs lies within 10% of the true measure of ancestry, as can be seen in Figure 1.

The small systematic errors in the estimation of ancestry with AIMs are likely due to the bounding of ancestry proportions at 0 and 1. The most a minor ancestral component can be underestimated is equal to its true value (for example, an ancestral estimate of 4% can at most be underestimated by 4%, if it is estimated to be 0%), but it can be overestimated much more substantially. Conversely, the major ancestral component cannot be overestimated by more than the difference between 100% and its true value, but it can be significantly underestimated. This effect is most notable in Figure 1 with African ancestry in Mexicans from Mexico City, where the bounding is visible as what appears to be a line with a slope of -1 that forms the lower limit of error estimates for ancestry proportions less than 0.05.

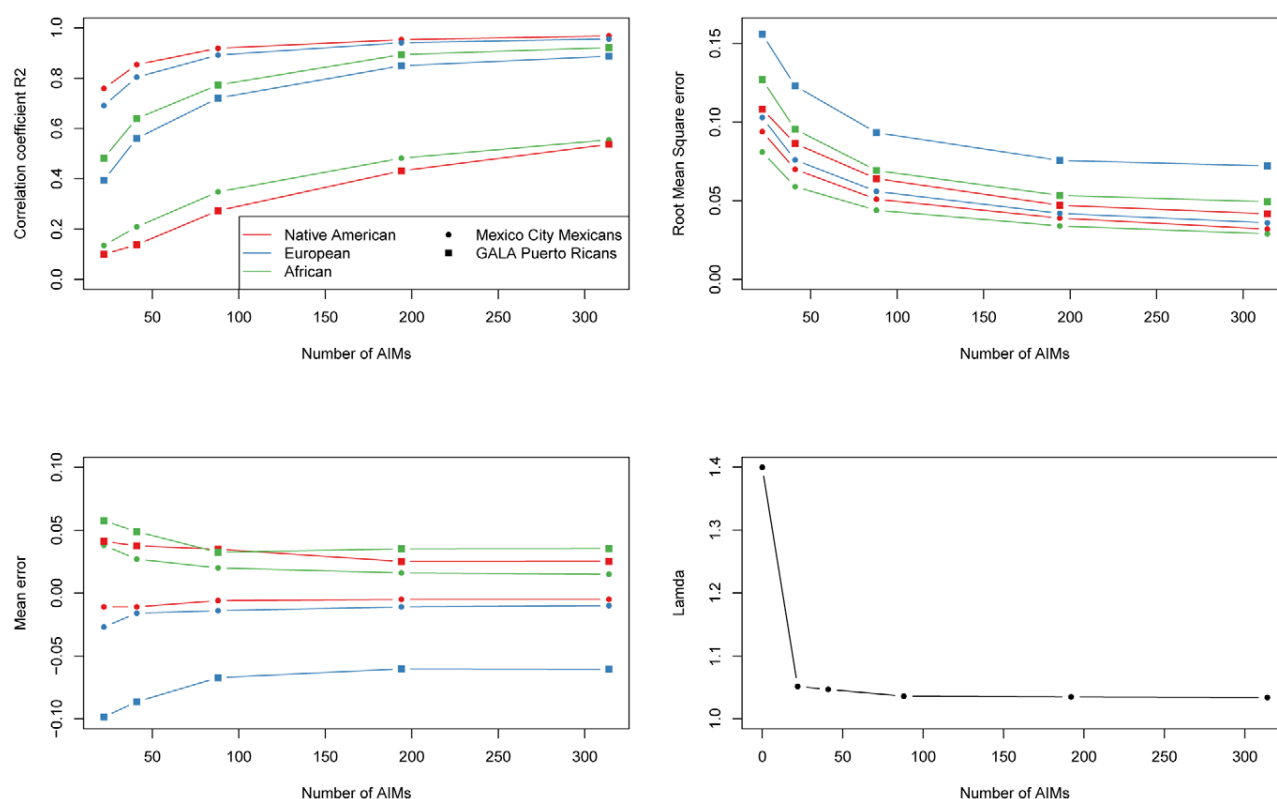


Figure 2. Performance of nested subsets of AIMs.
doi:10.1371/journal.pgen.1002554.g002

The slight increase in noise from AIMs panels compared to genome-wide estimates should then result in overestimates of minor components and underestimates of major components, consistent with observation.

We used the panel of AIMs to genotype 373 individuals from 18 Latin American populations. The samples were very diverse, and included individuals from several indigenous groups, African descendants and Mestizos from five different countries. Generally speaking there is strong concordance between ethnicity and admixture estimates. Specifically, seven out of eight indigenous samples showed a high degree of Native American ancestry. In particular, the four isolated groups from Venezuela (Warao, Panare and Pemón from the Amazon and Wayu from the northwestern region of Venezuela) showed very little evidence of European or African admixture. The three indigenous groups from Colombia (Coyaima, Pastos and Awa) had average Native American proportions higher than 80%, and a relatively small European contribution. That our AIMs panel could effectively estimate ancestry in lowland South American Native American populations (such as those in Venezuela) despite the fact that our AIMs were derived from Mesoamerican and Andean populations is reassuring and demonstrates that our strategy of excluding markers with significant heterogeneity ensures the generalizability of the markers. The indigenous Wichi from Argentina had considerably lower Native American ancestry and higher European ancestry (0.41 and 0.54, respectively) than the indigenous groups from Venezuela and Colombia. This is consistent with a recent study of Y-chromosomes that found widespread European paternal ancestry among Amerindian groups, including the Wichi, in Argentina [31]. Interestingly, we observed cryptic and previously unreported European admixture in the two isolated

Indigenous populations from Southern Colombia, a fairly common phenomenon in Native American populations [32].

In Bolivia, we found that the individuals from the Departments of Beni, Cochabamba and the Altiplano region of the La Paz Department had, on average, high Native American contributions. However, in the subtropical area of Yungas, many of the individuals recruited in the small community of Tocaña and one of the individuals recruited in the nearby town of Coroico had high African ancestry (median = 0.78, 0.74: 0.80). The subtropical Yungas region is home to several scattered Afro-Bolivian communities. These Afro-Bolivians are the descendants of African slaves who were brought to work on the Potosí mines and coca plantations [33]. Our data indicate that the admixture process in this Afro-Bolivian community has been primarily with the indigenous groups living in this region (median Native American ancestry = 0.13, 0.09: 0.20, median European ancestry = 0.04, 0.02: 0.06).

Two additional groups of African descent were included in this study, the Mulaló and Chocó from Colombia. African slaves were brought to Colombia early during the colonial period for gold mining, sugar cultivation, and cattle ranching. The proportion of African ancestry in these two Afro-Colombian groups was slightly lower than in the Afro-Bolivian community (0.54, 0.46: 0.69 in the Mulaló and 0.76, 0.64: 0.83 in the Chocó). Unlike the Afro-Bolivian sample from the Yungas region, in which most of the non-African contribution came primarily from indigenous groups, the two Afro-Colombian samples had similar European and Native American ancestral contributions (Table 4). This highlights the diverse history of admixture in different areas within Latin America. Similar observations have been reported by Castro de Guerra and colleagues [34,35], which compared two African

Table 4. Ancestry of Latin American populations.

Population	Country	Sample size	Native American Ancestry	European Ancestry	African Ancestry
Awa	Colombia (Southern)	22	0.80, 0.57: 0.87	0.17, 0.12: 0.37	0.02, 0.0: 0.05
Coyaima	Colombia (Central)	19	0.86, 0.83: 0.89	0.09, 0.07: 0.13	0.02, 0.01: 0.05
Pastos	Colombia (Southern)	36	0.83, 0.64: 0.87	0.16, 0.12: 0.31	0.02, 0.0: 0.04
Panare	Venezuela (Amazon)	20	1.00, 1.00: 1.00	0.0, 0.0: 0.0	0.0, 0.0: 0.0
Pemon	Venezuela (Amazon)	20	1.00, 1.00: 1.00	0.0, 0.0: 0.0	0.0, 0.0: 0.0
Warao	Venezuela (Amazon)	20	0.99, 0.97: 1.00	0.01, 0.0: 0.02	0.0, 0.0: 0.01
Wayu	Venezuela (North)	20	0.97, 0.84: 0.99	0.02, 0.0: 0.08	0.02, 0.0: 0.03
Wichi	Argentina	14	0.41, 0.12: 0.84	0.54, 0.13: 0.81	0.05, 0.01: 0.08
Maracaibo	Venezuela	20	0.28, 0.25: 0.36	0.60, 0.44: 0.62	0.12, 0.11: 0.15
Northern Chile	Chile	20	0.46, 0.37: 0.50	0.51, 0.43: 0.55	0.05, 0.03: 0.07
Southern Chile	Chile	20	0.51, 0.43: 0.55	0.45, 0.38: 0.53	0.06, 0.03: 0.08
Antioquia	Colombia	19	0.39, 0.35: 0.46	0.52, 0.48: 0.56	0.06, 0.04: 0.08
Antiplano	Bolivia	11	0.99, 0.98: 1.00	0.0, 0.0: 0.02	0.0, 0.0: 0.01
Chocó	Colombia	35	0.13, 0.10: 0.18	0.10, 0.07: 0.16	0.76, 0.64: 0.83
Mulaló	Colombia	28	0.18, 0.12: 0.26	0.25, 0.19: 0.20	0.54, 0.46: 0.69
Beni	Bolivia	10	0.94, 0.78: 0.96	0.04, 0.03: 0.22	0.01, 0.0: 0.03
Cochabamba	Bolivia	12	0.90, 0.86: 0.95	0.09, 0.05: 0.13	0.0, 0.0: 0.01
Yungas	Bolivia	27	0.25, 0.13: 0.97	0.03, 0.0: 0.05	0.70, 0.01: 0.82

Ancestries are given in median and 25th:75th percentiles.
doi:10.1371/journal.pgen.1002554.t004

derived populations in Venezuela and found that one, the Patenemos, showed mostly European ancestry, while the other population, Ganga, was principally admixed with Native American ancestry. We estimated that the time since admixture in the three samples of African descent is approximately 6 to 7 generations, corresponding to between 174 and 203 years, indicating that, the admixture process in these groups has been relatively recent. Though the point estimates of the years since admixture are approximately 50 to 100 years after the time when slaves were introduced into the region for gold and silver mining, because of the wide credible intervals, our estimates are not inconsistent with the historical record [36].

Our samples from the four Mestizo populations from Chile, Colombia, and Venezuela showed a wide variability in the ancestral proportions, though the primary ancestral contributions were European and Native American. Only some of the subjects from Maracaibo, on the Caribbean coast of Venezuela, had greater than 10% African ancestry, as did some of the Puerto Rican subjects used to validate the AIMs. This is unsurprising, given that the rest of our Mestizo populations are from Mexico, Chile and the Northwest of Colombia, areas where the slave trade was not prominent. This is consistent with the findings of Wang *et al.*, who examined thirteen Mestizo populations in Latin America and found extensive variation in Native American and European ancestry and relatively low levels of African ancestry [37]. We estimated between eight and thirteen generations since admixture for the mestizo samples, corresponding to between 230 and 375 years, reflecting the earlier settlement of substantial contingents of Europeans in Colombia than in Chile [38].

One striking finding in this paper is the rich ancestral variation in the Americas, even within a single country. For example, among the six Colombian populations examined (three Native American populations, one Mestizo population, and two Afro-Colombian populations), median Native American ancestry varied

between 0.13 in the Chocó and 0.86 in the Coyaima, African ancestry varied between 0.02 in the three Amerindian populations and 0.74 in the Chocó, and European ancestry varied between .09 in Coyaima and 0.52 in the Colombian mestizos. Likewise, even among the Bolivians in a single administrative department (state), there was a wide variation in African and Native American ancestry (Figure 3). These patterns of variation in ancestry within small regions seem to be a common feature across the Americas and have also been recently found in the island of Puerto Rico [3]. This has broad implications for genetic association studies in Latin American subjects, as there is a strong potential for population stratification, even in samples from a single country or a single administrative region within a country, and emphasizes the importance of incorporating ancestry estimates into future genetic association studies in these populations. We anticipate the primary use of this panel of AIMs will be to control for population stratification in genetic association and medical genetic studies. Thus, the ability of our panel of AIMs to effectively control for population stratification, as evidenced by its ability to reduce the genomic inflation factor in a highly stratified study of Type II diabetes in Mexican subjects, is an important source of validation. Even small subsets of AIMs from the panel adequately control for population stratification, suggesting that the panel should adequately cope with the significant patterns of variation in ancestry seen in Latin American. Nonetheless, because the panel of markers is not designed to identify within-continent heterogeneity, it is possible that it may not adequately control for finer population substructure.

In summary, we have developed and validated a panel of 446 AIMs to estimate European, Native American and African admixture proportions. The markers were selected to have low heterogeneity within continents, in order to be portable throughout the Americas. This panel was specifically designed to provide accurate individual admixture estimates and to control for the effects of population stratification in association studies in admixed

Ancestry Estimates of Latin American Populations

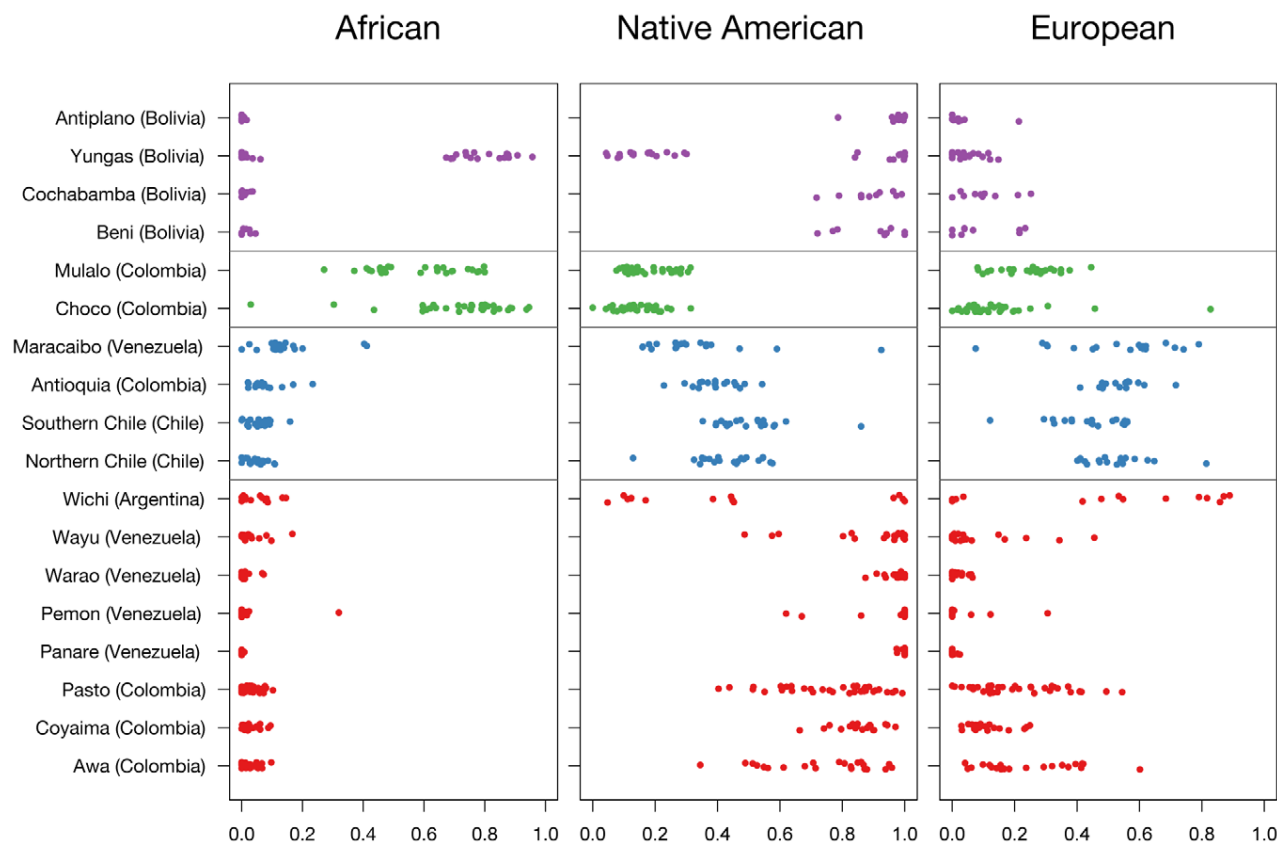


Figure 3. Ancestry estimates of Latin American populations.
doi:10.1371/journal.pgen.1002554.g003

populations. The use of this panel will minimize the risk of false positives in candidate gene studies, or in research efforts designed to replicate signals identified in genome-wide association studies, even in studies with substantial population stratification.

Our analysis of subsets of this panel has shown that to successfully control for population stratification in association studies, panels with 314, 194 and even 88 AIMs provide adequate estimates of the ancestral proportions with greatest variance that are strongly correlated with the genome-wide estimates (R^2 of 0.9 or higher) and have mean absolute error under 5%. Panels with 314, 194 and 88 AIMs all adequately controlled for the effects of population stratification in the Mexico City sample. The inflation factor (λ) was reduced from 1.40 when using sex and age as covariates, to less than 1.04 when incorporating ancestry estimates based on genome-wide data and panels of 314, 194 and 88 AIMs, and reasonable control for population stratification could be achieved with even smaller panels.

There are several important limitations to our AIMs panel. It is important to point out that the density of the markers in this panel is inadequate for admixture mapping, although the enclosed Python script could be used to identify a sufficient number of AIMs to perform an admixture mapping study [39]. Several research groups have already made available denser genome-wide panels of AIMs for admixture mapping in African Americans [40,41,42] and Hispanics [43,44,45], although none of these panels was designed for admixture models including three ancestral populations. The AIMs were selected for their

information content on African, European and Native American ancestry. These have been the major population groups contributing ancestry in the Americas since the 15th century. However, in many locations within the Americas, the history of human migration and admixture has been extremely complex, and has involved other population groups, such as East Asians and South Asians [46]. This panel of AIMs should be applied cautiously to populations (or individuals) with such complex admixture histories. Finally, while the panel has been validated to study the history of recent admixture in Latin America, it is unlikely to be effective in inferring finer scale population history.

As with all panels of AIMs, our panel is vulnerable to ascertainment bias, because the AIMs were selected to maximize the difference in continental ancestral allele frequencies. However, there are several factors that minimized the impact of this bias. First, we had a large sample size of all ancestral groups, particularly the European populations. Since the standard error of the estimate of allele frequency is inversely proportional to the square root of the number of individuals, the large sample sizes minimize the standard error in allele frequency estimates. Secondly, we used multiple populations within each continental group, and excluded any markers that showed large amounts of heterogeneity among ancestral groups within each continent. Thus, samples biased in one population (due to chance or genotyping error) are likely to have been filtered out. Finally, when we applied our panel to new populations, it produced

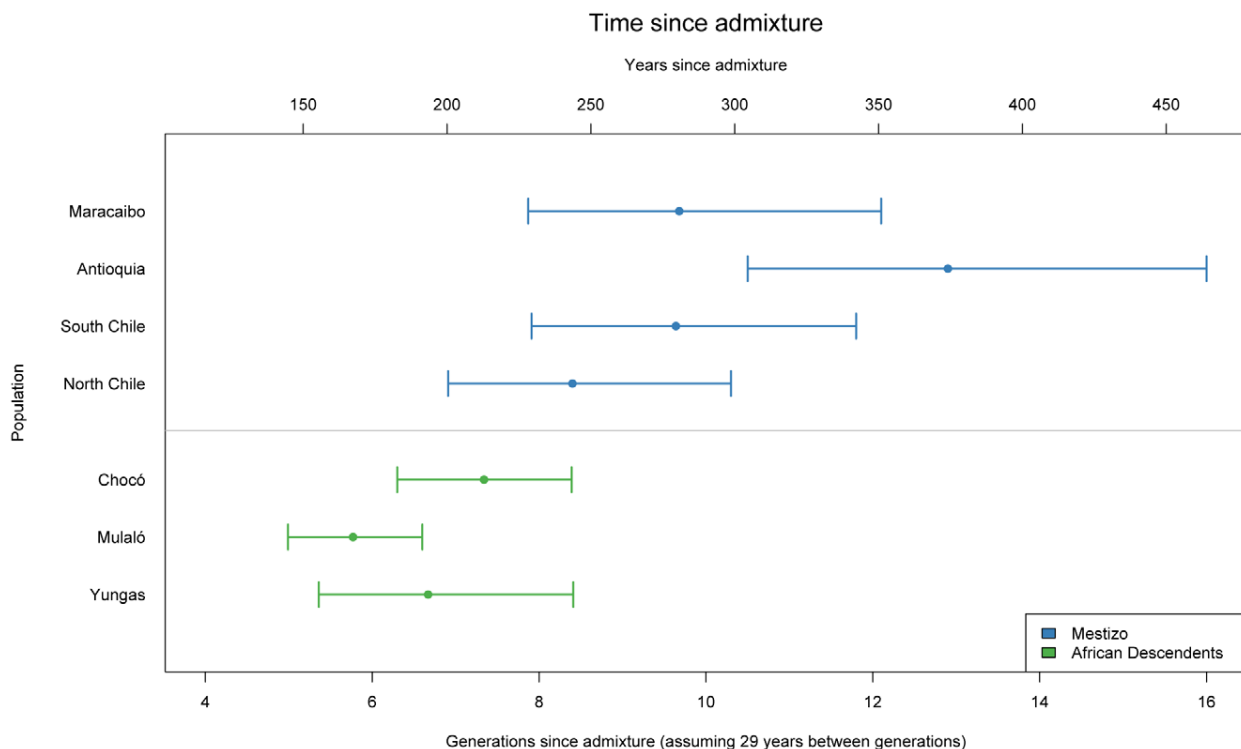


Figure 4. Time since admixture for Mestizo and African descendent populations.
doi:10.1371/journal.pgen.1002554.g004

credible ancestry estimates, which compare favorably to ancestry ascertained from genomewide data not subject to ascertainment bias.

This panel is intended to be an important resource for the community and we have provided both the source code for the algorithm to generate the AIMs, as well as allele frequency data and anonymized ancestral African, European, and shuffled Native American genotype information. We hope that investigators can use the selected panel of AIMs, which can be easily genotyped on readily available platforms, as a cost-effective tool to estimate continental ancestry in modern populations of the Americas.

Materials and Methods

Ethics statement

Informed consent was obtained for all subjects in all phases of this study, with input from local communities. These studies were approved by local institutional review boards and the relevant offices at each institution contributing samples (detailed information on approvals and consents for all samples available in Text S1).

Ancestral samples and genotyping

Subjects representing the three main continental ancestral groups making up modern Latin American populations were

Table 5. Ancestral populations used for this study.

Population	Designation	Sample size	Platform(s)
Utah residents with ancestry from Northern and Western Europe (HapMap Phase III)	CEU	56	Affymetrix 6.0/Illumina 1M
Toscani in Italy (HapMap Phase III)	TSI	44	Affymetrix 6.0/Illumina 1M
Spaniards from Spain	SPAIN	619	Affymetrix 6.0
Yoruba in Ibadan, Nigeria (HapMap Phase III)	YRI	53	Affymetrix 6.0/Illumina 1M
Luhya in Webuye, Kenya (HapMap Phase III)	LWK	50	Affymetrix 6.0/Illumina 1M
Aymara from La Paz, Bolivia	AYMARA	25	Affymetrix 6.0
Quechua from cerro de Pasco, Peru	QUECHUA	24	Affymetrix 6.0
Nahua from Central Mexico	NAHUA	14	Affymetrix 6.0
Maya from Campeche, Mexico	MAYAS	25	Affymetrix 500K/Illumina 550K
Tepehuano from Durango, Mexico	TEPHUANOS	22	Affymetrix 500K/Illumina 550K
Zapoteca from Oaxaca, Mexico	ZAPOTECAS	21	Affymetrix 500K/Illumina 550K

doi:10.1371/journal.pgen.1002554.t005

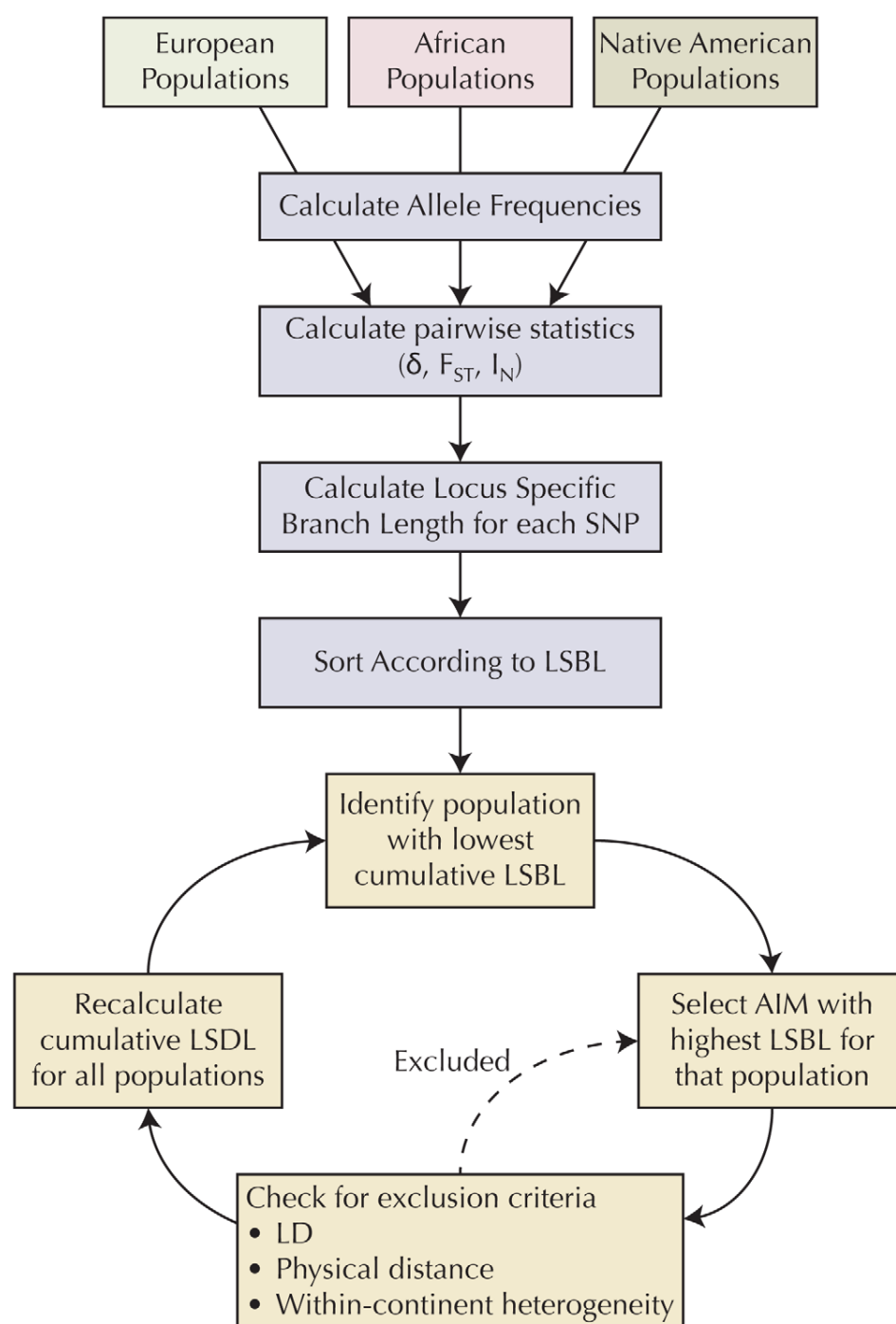


Figure 5. Algorithm for selecting AIMs.

doi:10.1371/journal.pgen.1002554.g005

obtained from a variety of sources. Hapmap Phase III genotype data for African and European populations was downloaded for this project, including West African (Yoruba in Ibadan, Nigeria, YRI) and East African (Luhya in Webuye, Kenya, LWK) as well as Northern European (Utah residents with ancestry from Northern and Western Europe, CEU) and Southern European (Toscani in Italy, TSI) individuals [47,48]. For populations including parent/child trios or duos (CEU, YRI), only genotypes from the parents were used. In addition, known cryptically related individuals were

removed [49]. Genotyping data for Europeans was further supplemented by a cohort of 619 samples of Spanish individuals, genotyped on the Affymetrix SNP 6.0 platform.

One hundred and thirty-one Native American subjects, from Mesoamerica (Nahua from Central Mexico, $n = 14$, Zapotecas from Oaxaca, Mexico, $n = 21$, and Maya from Campeche, $n = 25$) [16,30,50], from the Sierra Madre Occidental region (Tepehuanos from Durango in Northern Mexico, $n = 22$) and South America (Aymara from La Paz, Bolivia, $n = 25$, and Quechua from Cerro

Table 6. Samples used for validation.

Population	Ethnicity	Sample size	Platform(s)
GALA	Mexican	668	Affymetrix 6.0
GALA	Puerto Rican	803	Affymetrix 6.0
MGDP-INMEGEN	Mexican	312	Affymetrix 500K+Illumina 550
Mexico City	Mexican	1310	Affymetrix 5.0

doi:10.1371/journal.pgen.1002554.t006

de Pasco, Peru, $n=24$) [14,51] were used to determine Native American allele frequencies. These populations were genotyped either on the Affymetrix SNP 6.0 or on two platforms, the Affymetrix SNP 500K and the Illumina 550.

A summary of the populations used for this study and genotyping platforms is given in Table 5. Although, additional Native American subjects with genomewide data are available from the Human Genome Diversity Panel, these subjects were genotyped on the Illumina HumanHap 650k, and the intersection with the genotyping platforms used in our samples would have left fewer markers to be evaluated.

Quality control

Four major quality control tests were performed on the data using the program plink [52]. Individuals were excluded if they had greater than 10% missing alleles, if they were known to be related, or showed cryptic relatedness. For Native American populations, pairwise individuals were considered to have cryptic relatedness if their IBS scores showed a $Z1 > 0.15$ or a $Z2 > 0.03$ or if they had a proportion IBD (π) > 0.08 [53]. Europeans and African individuals were considered cryptically related if they had a $Z1 > 0.03$ or $Z2 > 0.03$, or if they had a proportion IBD > 0.03 . SNPs were included if the genotyping rate was greater than 90%

and excluded if they failed a χ^2 test for Hardy-Weinberg equilibrium at a significance threshold of 10^{-5} .

Stage one: AIM selection

Markers representing the intersection of the genotyping platforms used to genotype the ancestral populations, which met quality control criteria ($n=319,665$) were used as a basis for selecting AIMs. Figure 5 summarizes the methodology used to select AIMs. For each SNP for each ancestral group, allele frequency was calculated with the program plink [52]. For each marker, statistics of informativeness, including delta, F_{st} [54], and Rosenberg's informativeness for assignment statistic I_n [55] were calculated between each pair of ancestral populations (African/European, European/Native American, and African/Native American) based on reference allele frequencies. Locus specific branch length (LSBL) [56] statistics were created for each population and each statistic of informativeness to translate the pairwise metrics into a population-specific statistic. A balanced set of AIMs was selected by ensuring that the cumulative LSBL for each population was approximately equal. At each stage, we selected the polymorphism with the highest LSBL for the population with the lowest cumulative LSBL that met the inclusion criteria. Polymorphisms were excluded if they were in linkage disequilibrium ($r^2 \geq 0.1$) or within a predefined physical distance (≤ 500 kb pairs) of previously selected AIMs. This ensures maximum independent informativeness and that the AIMs were well distributed throughout the genome. In addition, in order for potential AIMs to be applicable to all subpopulations within a continental group, potential AIMs were also excluded if there was evidence of significant allele frequency heterogeneity between the samples representing each ancestral group (χ^2 p-value < 0.01). A script in the Python programming language that implements this algorithm and ancestral population allele frequency data are available for download.

Table 7. Latin American populations genotyped in stage III of this study.

Population	Country	Ethnicity	Sample size
Awa	Colombia (Southern)	Indigenous	22
Coyaima	Colombia (Central)	Indigenous	19
Pastos	Colombia (Southern)	Indigenous	36
Panare	Venezuela (Amazon)	Indigenous	20
Pemon	Venezuela (Amazon)	Indigenous	20
Warao	Venezuela (Amazon)	Indigenous	20
Wayu	Venezuela (North)	Indigenous	20
Wichi	Argentina	Indigenous	14
Maracaibo	Venezuela	Mestizo (admixed)	20
Northern Chile	Chile	Mestizo (admixed)	20
Southern Chile	Chile	Mestizo (admixed)	20
Antioquia	Colombia	Mestizo (admixed)	19
Antiplano	Bolivia	Mestizo (admixed)	11
Chocó	Colombia	Afro-Colombian	35
Mulaló	Colombia	Afro-Colombian	28
Beni	Bolivia	Multi-ethnic (Mestizo and Indigenous)	10
Cochabamba	Bolivia	Multi-ethnic (Mestizo and Indigenous)	12
Yungas	Bolivia	Multi-ethnic (Indigenous, Afro-Bolivian)	27

doi:10.1371/journal.pgen.1002554.t007



Figure 6. Origin of samples used in this study. Labels in purple correspond to the Native American ancestral populations, labels in red to the validation samples, and labels in black to the 18 populations from throughout the Americas. MGD-P-INMEGEN samples were collected throughout Mexico (see Figure S1). GALA Mexico samples were also collected in the San Francisco Bay Area, CA. GALA Puerto Rico samples were also collected in New York, NY.
doi:10.1371/journal.pgen.1002554.g006

Stage two: Validation of the panel of AIMs

In order to validate the panel of AIMs, estimates of ancestry using the panel were compared to estimates of ancestry using genome-wide data. Four admixed samples were used for validation. The first two datasets were parents of Puerto Rican and Mexican subjects with asthma genotyped on the Affymetrix 6.0 GeneChip as part of the Genetics of Asthma in Latino Americans (GALA) study [20]. The third sample consists of 1,310 individuals from Mexico City participating in a type 2 diabetes study that were genotyped with the Affymetrix 5.0 GeneChip. The fourth sample contains 312 subjects in the Mexican Genome Diversity Project (MGDP) recruited by the National Institute of Genomic Medicine (INMEGEN) from throughout Mexico, including 48 subjects from Guanajuato, 50 subjects from Guerrero, 48 subjects from Sonora, 17 subjects from Tamaulipas, 50 subjects from Veracruz, 49 subjects from Yucatan, and 50 subjects from Zacatecas [30,50]. A map of the geographic distribution of MGDP-INMEGEN samples is shown in Figure S2. A description of all the validation samples is shown in Table 6.

We implemented a three-population model to estimate individual ancestry proportions from genome-wide data using the program ADMIXTURE [57]. We filtered our genome-wide markers to eliminate markers in linkage disequilibrium at $r^2 > 0.8$. Genotypes from ancestral populations described above defined the ancestral clusters relevant to Latin Americans. We also estimated ancestry using the panel of AIMs identified with the protocol above. The performance of the AIMs panel was established by calculating the correlation coefficient (R^2) and measures of discordance (mean error, mean absolute error, and root mean squared error).

Stage three: Genotyping of populations throughout Latin America

Using the validated AIMs panel, we genotyped 18 populations collected from Bolivia, Colombia, Venezuela, Argentina, and Chile. A description of the origin of the samples is provided in Text S1 and in Table 7 and Figure 6.

Genotyping

Subjects were genotyped on a Sequenom platform with the 400 most informative AIMs identified in phase I. AIMs were included in the final analysis if they had a genotyping call rate greater than 95% and Hardy-Weinberg equilibrium in each population individually. We required that all samples had genotyping missing data rates of $<10\%$. Sample population groups were excluded if they have average genotyping data rates of $<10\%$.

Software and statistical analysis

File merging, strand flipping, allele frequency determination, linkage disequilibrium calculations, and identity by descent estimations were performed with the program plink [52]. The algorithm to develop the panel of AIMs was implemented in Python version 2.6 [58]. Individual ancestral estimates were performed with a three-population model using a model-based likelihood estimation using the program ADMIXTURE [57]. Statistical analyses were performed with R and Python [58,59]. Estimation of time since admixture was performed using the program ADMIXMAP [60,61,62], assuming an average of 29 years per generation [63].

References

1. Wang S, Ray N, Rojas W, Parra MV, Bedoya G, et al. (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* 4: e1000037. doi:10.1371/journal.pgen.1000037.
2. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, et al. (2010) Colloquium paper: genome-wide patterns of population structure and admixture

Web resources

Source code for the AIMs selection script is available at <http://bts.ucsf.edu/burchard/>.

Supporting Information

Figure S1 QQ plots of genetic association studies of Diabetes in Mexicans, using nested sets of AIMs.
(PDF)

Figure S2 Origin of Mexican samples from MGDP-INMEGEN. Locations in purple correspond to Native American populations; those in red correspond to admixed MGDP-INMEGEN populations used for validation.
(PDF)

Table S1 AIMs used.
(DOCX)

Text S1 Supplemental methods, including detailed ethics statement and description of samples obtained for genotyping of populations throughout the Americas.
(DOCX)

Acknowledgments

The authors would like to thank several individuals and institutions for their contributions. For samples collected in Bolivia: Hospital Clínico Viedma, Hospital del Niño M.A.V., Hospital Materno Infantil Germán Urquidí, Departamento de Cochabamba; Hospital de Clínicas de la ciudad de La Paz, Hospital Municipal de Copacabana, Hospital Universitario de Coroico y Tocaña community, Departamento de La Paz; Hospital Germán Bush, Departamento del Beni. For samples collected in Venezuela: Miguel Ángel Ciurillo S. M.D., Ph.D., Universidad Centroccidental “Lisandro Alvarado,” Lara State. For samples collected in Colombia: Hugo German Burgos Figueroa (Cruz Roja Ecuatoriana). For samples collected in Argentina: Ulises Toscanini, PRICAI-Fundación Favaloro, Buenos Aires. The authors would also like to thank Leticia Sebastian Medina and Fabiola Morales Mandujano from the INMEGEN Genotyping and Expression Analysis Illumina Unit and Gisela Ortiz Ramos from the INMEGEN Genotyping and Expression Analysis Affymetrix Unit. The authors are grateful to Jeffrey M. Drazen, M.D.; Scott Weiss, M.D.; Ed Silverman, M.D., Ph.D.; and Homer A. Boushey, M.D., for all of their effort towards the creation of the GALA Study. The authors would like to acknowledge the families and the subjects in all studies for their participation. The authors would also like to thank the numerous health care providers, community clinics, and local community groups for their support and participation. Some computations were performed using the UCSF Biostatistics High Performance Computing System.

Author Contributions

Conceived and designed the experiments: JMG JCF-L CRG DS CB EZ EGB RH EP AC. Performed the experiments: JMG JCF-L CRG CF-R LUF GJ-S MT CRP YR AS EN CE EP AC. Analyzed the data: JMG JCF-L CRG EP AC. Contributed reagents/materials/analysis tools: JCF-L CF-R AH-M AVC LUF PR GJ-S ISZ MT CRP YR AS LB WZ GB FRG AI PT LP FM AB GG TB FL-V LGM EV MC JE JR-S WR-C RC JGF MS EGB EP AC. Wrote the paper: JMG JCF-L CRG JB-S CF-R MV GJ-S EN DS CB EZ EGB RH EP AC.

among Hispanic/Latino populations. *Proc Natl Acad Sci U S A* 107 Suppl 2: 8954–8961.

3. Via M, Gignoux CR, Roth LA, Fejerman L, Galanter J, et al. (2011) History shaped the geographic distribution of genomic admixture on the island of Puerto Rico. *PLoS ONE* 6: e16513. doi:10.1371/journal.pone.0016513.

4. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, et al. (1998) Estimating African American admixture proportions by use of population-specific alleles. *American journal of human genetics* 63: 1839–1851.
5. Phillips C, Prieto L, Fondevila M, Salas A, Gomez-Tato A, et al. (2009) Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS ONE* 4: e6583. doi:10.1371/journal.pone.0006583.
6. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.
7. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. *American journal of human genetics* 74: 965–978.
8. Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, et al. (2010) Genetic ancestry in lung-function predictions. *The New England journal of medicine* 363: 321–330.
9. Julian CG, Wilson MJ, Lopez M, Yamashiro H, Tellez W, et al. (2009) Augmented uterine artery blood flow and oxygen delivery protect Andeans from altitude-associated reductions in fetal growth. *American journal of physiology Regulatory, integrative and comparative physiology* 296: R1564–1575.
10. Allison MA, Peralta CA, Wassel CL, Aboyans V, Arnett DK, et al. (2010) Genetic ancestry and lower extremity peripheral artery disease in the Multi-Ethnic Study of Atherosclerosis. *Vascular medicine* 15: 351–359.
11. Fejerman L, Romieu I, John EM, Lazcano-Ponce E, Huntsman S, et al. (2010) European ancestry is positively associated with breast cancer risk in Mexican women. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 19: 1074–1082.
12. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American journal of human genetics* 65: 220–228.
13. Humes KR, Jones NA, Ramirez RR, United States. Bureau of the Census. (2011) Overview of race and Hispanic origin : 2010. Washington, D.C.: U.S. Dept. of Commerce, Economics and Statistics Administration, U.S. Census Bureau 23 p.
14. Brutsaert TD, Parra EJ, Shriver MD, Gamboa A, Palacios JA, et al. (2003) Spanish genetic admixture is associated with larger V(O₂) max decrement from sea level to 4338 m in Peruvian Quechua. *Journal of applied physiology* 95: 519–528.
15. Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, et al. (2004) Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. *Human genetics* 114: 263–271.
16. Bonilla C, Gutierrez G, Parra EJ, Kline C, Shriver MD (2005) Admixture analysis of a rural population of the state of Guerrero, Mexico. *American journal of physical anthropology* 128: 861–869.
17. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, et al. (2006) Population stratification confounds genetic association studies among Latinos. *Human genetics* 118: 652–664.
18. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998–1003.
19. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* 102: 15942–15947.
20. Burchard EG, Avila PC, Nazario S, Casal J, Torres A, et al. (2004) Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am J Respir Crit Care Med* 169: 386–392.
21. Below JE, Gamazon ER, Morrison JV, Konkashbaev A, Pluzhnikov A, et al. (2011) Genome-wide association and meta-analysis in populations from Starr County, Texas, and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals. *Diabetologia* 54: 2047–2055.
22. Parra EJ, Below JE, Krithika S, Valladares A, Barta JL, et al. (2011) Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas. *Diabetologia* 54: 2038–2046.
23. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55: 997–1004.
24. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324: 1035–1044.
25. Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J (2011) A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. *Eur J Hum Genet* 19: 84–88.
26. Lovejoy PE (2000) Transformations in slavery : a history of slavery in Africa. Cambridge, UK; New York: Cambridge University Press. xxii, 367 p.
27. Curtin PD (1969) The Atlantic slave trade; a census. Madison: University of Wisconsin Press. xix, 338 p.
28. Eltis D, Richardson D (2010) Atlas of the transatlantic slave trade. The Lewis Walpole series in eighteenth-century culture and history. New Haven, Conn.: Yale University Press.
29. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, et al. (2010) Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* 107 Suppl 2: 8954–8961.
30. Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, et al. (2009) Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proceedings of the National Academy of Sciences of the United States of America* 106: 8611–8616.
31. Diaz-Lacava A, Walier M, Penacino G, Wienker TF, Baur MP (2011) Spatial assessment of Argentinean genetic admixture with geographical information systems. *Forensic science international Genetics* 5: 297–302.
32. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, et al. (2007) Genetic variation and population structure in native Americans. *PLoS Genet* 3: e185. doi:10.1371/journal.pgen.0030185.
33. Busdiecker S (2009) Where Blackness Resides: Afro-Bolivians and the Spatializing and Racializing of the African Diaspora. *Radical History Review* 2009: 105–116.
34. Castro de Guerra D, Arvelo H, Pinto-Cisternas J (1993) Estructura de población y factores influyentes en dos pueblos negros venezolanos. *América Negra* 5: 37–47.
35. Bortolini MC, Salzano FM, de Azevedo Weimer T (1995) Inter and intrapopulation genetic diversity in Afro-Venezuelan and African populations. *Interciencia* 20: 90–93.
36. Klein HS (1988) African Slavery in Latin America and the Caribbean. New York: Oxford University Press.
37. Wang S, Ray N, Rojas W, Parra MV, Bedoya G, et al. (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* 4: e1000037. doi:10.1371/journal.pgen.1000037.
38. Sánchez-Albornoz N (1977) La Población de América Latina: Desde los Tiempos Precolombinos al Año 2000. Madrid: Alianza Editorial.
39. Winkler CA, Nelson GW, Smith MW (2010) Admixture mapping comes of age. *Annual review of genomics and human genetics* 11: 65–89.
40. Chen G, Shriner D, Zhou J, Doumatey A, Huang H, et al. (2010) Development of admixture mapping panels for African Americans from commercial high-density SNP arrays. *BMC genomics* 11: 417.
41. Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, et al. (2006) A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *American journal of human genetics* 79: 640–649.
42. Smith MW, Patterson N, Lautenberger JA, Truvelo AL, McDonald GJ, et al. (2004) A high-density admixture map for disease gene discovery in African Americans. *American journal of human genetics* 74: 1001–1013.
43. Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, et al. (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. *American journal of human genetics* 80: 1171–1178.
44. Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, et al. (2007) A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *American journal of human genetics* 80: 1014–1023.
45. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, et al. (2007) A genomewide admixture map for Latino populations. *American journal of human genetics* 80: 1024–1036.
46. Salzano FM, Bortolini MC (2002) The Evolution and Genetics of Latin American Populations. Cambridge: Cambridge University Press.
47. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.
48. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52–58.
49. Pemberton TJ, Wang C, Li JZ, Rosenberg NA (2010) Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *American journal of human genetics* 87: 457–464.
50. INMEGEN Mexican Genome Diversity Project (MGDP).
51. Vargas M, Vargas E, Julian CG, Armaza JF, Rodriguez A, et al. (2007) Determinants of blood oxygenation during pregnancy in Andean and European residents of high Altitude. *Am J Physiol Regul Integr Comp Physiol* 293: 1303–1312.
52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.
53. Weir BS, Anderson AD, Hepler AB (2006) Genetic relatedness analysis: modern data and new challenges. *Nature reviews Genetics* 7: 771–780.
54. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews Genetics* 10: 639–650.
55. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *American journal of human genetics* 73: 1402–1422.
56. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human genomics* 1: 274–286.

Ancestry Informative Markers Panel Development

57. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19: 1655–1664.
58. Rossum Gv, Boer Jd (1991) Interactively testing remote servers using the Python programming language *CWI Quarterly* 4: 283–304.
59. Team RDC (2010) R: A Language and Environment for Statistical Computing. In: *Computing RFIIS*, ed. 2.12.1 ed. Vienna, Austria.
60. McKeigue PM, Carpenter JR, Parra EJ, Shriver MD (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annals of human genetics* 64: 171–186.
61. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, et al. (2003) Control of confounding of genetic associations in stratified populations. *American journal of human genetics* 72: 1492–1504.
62. Montana G, Hoggart C (2007) Statistical software for gene mapping by admixture linkage disequilibrium. *Briefings in bioinformatics* 8: 393–395.
63. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American journal of physical anthropology* 128: 415–423.

Bloque 1.

***V.3. A Melting pot of multicontinental
mtDNA lineages in admixed venezuelans***

Alberto Gómez-Carballa, Ana Ignacio-Veiga, Vanesa Álvarez-Iglesias, Ana Pastoriza-Mourelle, Yarimar Ruiz, Lennie Pineda, Ángel Carracedo, and Antonio Salas.

(American Journal of Physical Anthropology, 2012, 147:78-87)

A Melting Pot of Multicontinental mtDNA Lineages in Admixed Venezuelans

Alberto Gómez-Carballa,¹ Ana Ignacio-Veiga,¹ Vanesa Álvarez-Iglesias,¹ Ana Pastoriza-Mourelle,¹ Yarimar Ruiz,¹ Lennie Pineda,^{1,2} Ángel Carracedo,¹ and Antonio Salas^{1*}

¹Unidade de Xenética, Instituto de Medicina Legal and Departamento de Anatomía Patolóxica y Ciencias Forenses, Facultade de Medicina, 15782, Universidade de Santiago de Compostela, Galicia, Spain

²Unidad de Genética Médica, Facultad de Medicina, Universidad de Zulia, Maracaibo, Zulia, Venezuela

KEY WORDS entire genomes; coding region; HVS-I/II; SNPs; haplogroup; phylogeny

ABSTRACT The arrival of Europeans in Colonial and post-Colonial times coupled with the forced introduction of sub-Saharan Africans have dramatically changed the genetic background of Venezuela. The main aim of the present study was to evaluate, through the study of mitochondrial DNA (mtDNA) variation, the extent of admixture and the characterization of the most likely continental ancestral sources of present-day urban Venezuelans. We analyzed two admixed populations that have experienced different demographic histories, namely, Caracas ($n = 131$) and Pueblo Llano ($n = 219$). The native American component of admixed Venezuelans accounted for 80% (46% haplogroup [hg] A2, 7% hg B2, 21% hg C1, and 6% hg D1) of all mtDNAs; while the sub-Saharan and European contributions made up ~10% each, indicating that Trans-Atlantic immigrants have only partially erased the native

American nature of Venezuelans. A Bayesian-based model allowed the different contributions of European countries to admixed Venezuelans to be disentangled (Spain: ~38.4%, Portugal: ~35.5%, Italy: ~27.0%), in good agreement with the documented history. Seventeen entire mtDNA genomes were sequenced, which allowed five new native American branches to be discovered. B2j and B2k, are supported by two different haplotypes and control region data, and their coalescence ages are 3.9 k.y. (95% C.I. 0–7.8) and 2.6 k.y. (95% C.I. 0.1–5.2), respectively. The other clades were exclusively observed in Pueblo Llano and they show the fingerprint of strong recent genetic drift coupled with severe historical consanguinity episodes that might explain the high prevalence of certain Mendelian and complex multifactorial diseases in this region. *Am J Phys Anthropol* 147:78–87, 2012. © 2011 Wiley Periodicals, Inc.

An ancient watering hole near the coast of western Venezuela, known as Taima-Taima (a mastodon killing/butchering site), became one of the most popular and significant archaeological findings of the mid-twentieth century. It yielded evidence of humans in northern South America during the terminal Pleistocene-early Holocene periods (14,000–10,000 B.P.) (<http://www.bradshawfoundation.com/>; (Ochsenius and Gruhn, 1979; Dillehay, 2000). It is difficult to estimate how many indigenous people lived in Venezuela before Colonial times. Morón et al. (1994) estimated this figure as being about 300,000 during the initial years of the 16th century. The demographic scenario of Venezuela, as in any other country in America, changed dramatically with the arrival of Europeans and the Trans-Atlantic African slave trade. According to the census, the indigenous populations decayed as follows: 103,492 inhabitants (year 1936); 100,600 (1941); 98,687 (1950), and 75,604 (1961), and today there are only about 60,000 indigenous inhabitants left (Morón, 1994). The Comisión Indigenista from Venezuela elevated this number to around 150,000 (year 1977), and the Ethnologue catalogue (Lewis, 2009) refers to a similar figure of 145,230 (which makes up ~0.5% of the Venezuelan population). Natives form minor ethnic groups and they have preserved about 40 living languages (Lewis, 2009). Some of these native populations are in danger of becoming extinct, together with their languages and dialects, because of their small population sizes. Today, the most representative indigenous populations are the Wayuu and the Waraos.

Immigration from Spain in colonial times coupled with the forced introduction of African slaves into Venezuela lead to a dramatic reduction of the native American population in recent times. The contribution of Europe was particularly important immediately before and after the Second World War (Table 1). At that time, the population of the country was only about 5 million people. Massive waves of Italians (~300,000), predominantly from Southern Italy (e.g., Sicily, Campania, Abruzzo, Apulia), arrived at La Guaira harbor (today the capital city of the Venezuelan state of Vargas), constituting the main

Additional Supporting Information may be found in the online version of this article.

Alberto Gómez-Carballa, Ana Ignacio-Veiga, and Antonio Salas contributed equally to this work.

Grant sponsor: Ministerio de Ciencia e Innovación; Grant number: SAF2008-02971. Grant sponsor: Fundación de Investigación Médica Mutua Madrileña; Grant number: 2008/CL444.

*Correspondence to: Antonio Salas, Unidade de Xenética Forense, Departamento de Anatomía Patolóxica e Ciencias Forenses, Universidade de Santiago de Compostela, Galicia, Spain. E-mail: antonio.salas@usc.es

Received 25 May 2011; accepted 9 September 2011

DOI 10.1002/ajpa.21629

Published online 25 November 2011 in Wiley Online Library (wileyonlinelibrary.com).

TABLE 1. Adapted from the census of immigrants in Venezuela according to the INE

Census period	Europe	Spain	Portugal	Italy	Spain +Portugal +Italy	America	Colombia
Until 1939	1,656	726 (44%)	168 (10%)	370 (22%)	1,264 (76%)	3,650	2,742 (75%)
1940–1969	113,796	48,771 (43%)	21,731 (19%)	35,978 (32%)	106,480 (94%)	111,654	95,772 (86%)
1970–1979	26,583	6,865 (26%)	14,526 (55%)	3,399 (13%)	24,790 (93%)	208,084	169,419 (81%)
1980–1999	17,584	4,050 (23%)	7,259 (41%)	3,399 (19%)	11,063 (63%)	270,758	207,686 (77%)
2000–	1,208	280 (23%)	63 (5%)	224 (19%)	567 (47%)	41,144	35,188 (86%)
Not-declared	36,561	15,956 (44%)	9,730 (27%)	224 (1%)	17,153 (47%)	131,152	98,389 (75%)
Total	197,388	76,648 (39%)	53,478 (27%)	49,338 (25%)	179,464 (91%)	766,441	609,196 (79%)

The percentages (in brackets) indicate the relative proportions of the three main European countries (Spain, Portugal, and Italy) in the total European contribution, and the proportion of Colombia in the total American contribution.

European colony in Venezuela. Immigration from Portugal, from the 1940s to the 1980s, created the second most important Portuguese colony in Latin America (after Brazil) (Table 1). The Canary Islands, followed by Galicia, a region located in the northwestern corner of Spain, were the main Spanish contributors to Venezuela during colonial times and in the Second World War. Today, Venezuela has the second largest Spanish community in America, right after Argentina (INE; Instituto Nacional de Estadística, República Bolivariana de Venezuela, INE; <http://www.ine.gov.ve/>).

Mitochondrial DNA (mtDNA) is one of the most popular genetic markers used by molecular anthropologists for the reconstruction of past demographic events and for unraveling admixture processes that have occurred during the recent history of human populations; including those occurring in America, some examples are (Tamm et al., 2007; Achilli et al., 2008; Fagundes et al., 2008; Perego et al., 2009; Sandoval et al., 2009; Perego et al., 2010). A study by Merriwether et al. (2000) was the first attempt to analyze the mtDNA variation of an indigenous population from Venezuela (and Brazil), the Yanomami, a classic case study in anthropology and population genetics. This population lives in the Amazonian rainforest; the analysis by Merriwether et al. indicated that about 91% of the mtDNA belonged to haplogroup (hgs) D (40%) and C (51%). The observation of a major native American component in Yanomami was also described in a later study by Williams et al. (2002) (referred to as Yanomamö in their study); however, this time a different hg pattern was found: 56% hg B, 32% hg C, and 12% hg D. The study of Vona et al. (2005) focused on the Guahibo ethnic group and also showed a remarkable native American nature of their mtDNA, with hg frequencies of ~47% for A2, ~49% for C1, and ~3% for B1. Using restriction fragment length polymorphism (RFLP) analysis, Martínez et al. (2007) described the hg frequencies in Caracas, indicating a higher proportion of native American hgs (ranging from ~43 to ~72%, depending on the socio-economic level of the population). More recently, Lander et al. (2008) analyzed a sample from Caracas with the main focus on forensic genetics by reporting the profiles that covered most of the control region.

The main aim of this study was to analyze the admixture processes that have occurred in Venezuela over the last few centuries, where the native American component of the country has been progressively erased in urban Venezuelans by the arrival of immigrants coming from Europe and slaves brought from sub-Saharan Africa, as well as in more recent times, by immigrants coming from neighboring American countries. Two different communities were sampled for this study: one of them was in the city of Caracas, and was analyzed

together with another community in this city reported by Lander et al. (2008), while a second community was sampled in the city of Pueblo Llano and its surroundings (Estado Mérida).

MATERIAL AND METHODS

Samples

A total of 31 samples were collected from the city of Caracas (Venezuela) and 219 from the district of Pueblo Llano, located in Estado Mérida (Venezuela). Written informed consent was required for all samples. The sample from Caracas was merged with another one reported in the literature also collected in this city (Lander et al., 2008) and analyzed together; setting the total sample size to 131 individuals. Maternally related individuals were excluded from the analysis at the time of sample collection. Donors sharing surnames were interviewed in order to avoid close relatives.

The anonymity of the samples was preserved at the time of collection. The DNA extracts were submitted to the laboratory in Santiago de Compostela (Galicia; Spain), where the genotyping was carried out. In addition, the study was approved by the Ethical committee of the University of Santiago de Compostela. The study conformed to the Spanish Law for Biomedical Research (Law 14/2007-3 of July).

Genotyping protocols and nomenclature

All of the samples were PCR amplified and sequenced for the whole control region following the protocols described in Álvarez-Iglesias et al. (2009). The samples were allocated to main hgs based on control region information. Those belonging to Eurasian hgs were genotyped for a total of 71 mtDNA Single Nucleotide Polymorphisms (mtSNPs) defining the main European branches and internal variations of hg R0, following Álvarez-Iglesias et al. (2009) (Supporting Information Table S1). Furthermore, samples belonging to native American hgs were genotyped for 30 mtSNPs defining the main and minor branches of the native American phylogeny, as performed by Álvarez-Iglesias et al. (2007) (Supporting Information Table S1).

Seventeen native American samples were sequenced for the complete mtDNA genome, following the protocols described in Brisighelli et al. (2009) and Cerezo et al. (2011).

Mitochondrial DNA variation is referred with respect to the revised Cambridge Reference Sequence or rCRS (Andrews et al., 1999) (Supporting Information Table S1). The hg nomenclature is based on Phylotree (<http://www.phylotree.org/>) (Van Oven and Kayser, 2008). For the sake of simplicity, *L*-hg will denote in what follows hgs *L*

TABLE 2. Diversity indices in Venezuelan mtDNAs and continental variations in the main American and African regions. The indices were computed using the common segment of the HVSI region from position 16090 to 16365

	<i>n</i>	<i>S</i>	<i>H</i>	Π	<i>M</i>
Pueblo Llano (All)	199	55	0.935 \pm 0.010	0.020 \pm 0.001	5.5
Caracas (All)	131	73	0.958 \pm 0.012	0.026 \pm 0.001	7.0
Venezuela (All) ^a	330	89	0.949 \pm 0.000	0.024 \pm 0.001	6.5
Pueblo Llano: Native American component ^b	177	46	0.923 \pm 0.012	0.019 \pm 0.001	5.4
Caracas: Native American component ^b	84	42	0.901 \pm 0.025	0.021 \pm 0.001	5.7
Venezuela: Native American component ^{a,b}	261	58	0.919 \pm 0.012	0.020 \pm 0.001	5.5
North America ^b	2260	149	0.957 \pm 0.000	0.020 \pm 0.000	5.7
Meso-America ^b	1475	155	0.961 \pm 0.000	0.022 \pm 0.000	5.9
South America ^b	3082	170	0.962 \pm 0.000	0.021 \pm 0.000	5.6
Pueblo Llano: African component ^c	8	4	0.643 \pm 0.184	0.004 \pm 0.002	1.2
Caracas: African component ^c	26	39	0.988 \pm 0.014	0.030 \pm 0.014	8.2
Venezuela: African component ^{a,c}	34	44	0.986 \pm 0.011	0.031 \pm 0.002	8.4
West-Central Africa	6868	149	0.985 \pm 0.000	0.026 \pm 0.000	6.6
Southwest Africa	522	79	0.991 \pm 0.002	0.034 \pm 0.001	9.2
Southeast Africa	1118	85	0.970 \pm 0.003	0.031 \pm 0.000	8.5
East Africa	1054	134	0.995 \pm 0.000	0.031 \pm 0.000	8.2
South Africa	357	70	0.967 \pm 0.000	0.026 \pm 0.001	6.8
North Africa	3675	141	0.955 \pm 0.000	0.015 \pm 0.000	3.9

NOTE: *n* = sample size; *S* = number of segregating sites; *h* = haplotype diversity; π = nucleotide diversity; *M* = average number of pairwise differences (mismatch observed mean).

^a Caracas and Pueblo Llano are considered together.

^b Only haplotypes of Native American ancestry are considered.

^c Only haplotypes of recent sub-Saharan African ancestry (macro L-hg) are considered.

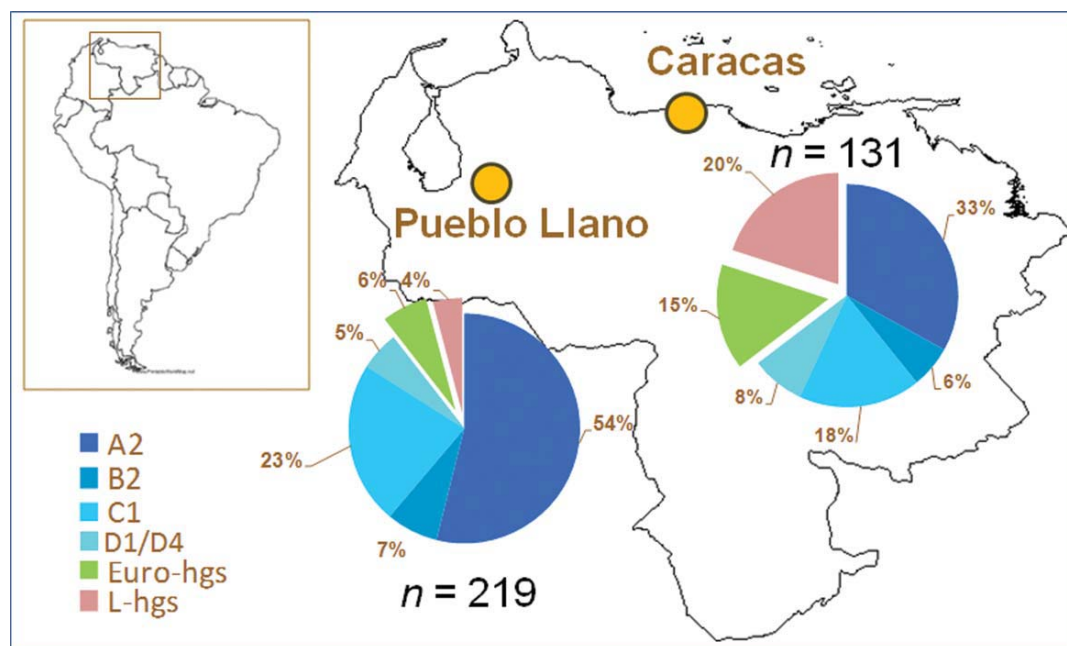


Fig. 1. Location of the two Venezuelan samples analyzed in the present study, Caracas and Pueblo Llano, and the frequency distribution of the main native American hgs (A2, B2, C1, and D1), the African L-hgs and the European hgs. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

excluding branches *M* and *N* (therefore referring to the most genuine sub-Saharan component).

Monitoring genotyping errors

We used the mtDNA tree as a reference to avoid errors in mtDNA profiles as much as possible (Bandelt et al., 2006). Unexpected mtSNP patterns were confirmed by repeating the SNP genotyping using single-plex mini-sequencing and automatic sequencing, as performed in Álvarez-Iglesias et al. (2007).

Statistical analysis and databases

The DnaSP 4.10.3 software (Rozas et al., 2003) was used for the computation of different diversity indices, including haplotype (*H*) and nucleotide (π) diversities and the mean number of pairwise differences (*M*) (Tajima, 1983; Nei, 1987; Tajima, 1993) (Table 2).

Diversity indices and interpopulation comparisons (e.g. Fig. 1) were carried out using the sequence range 16090 to 16365, since this is the common segment reported in the literature. Moreover, the problematic

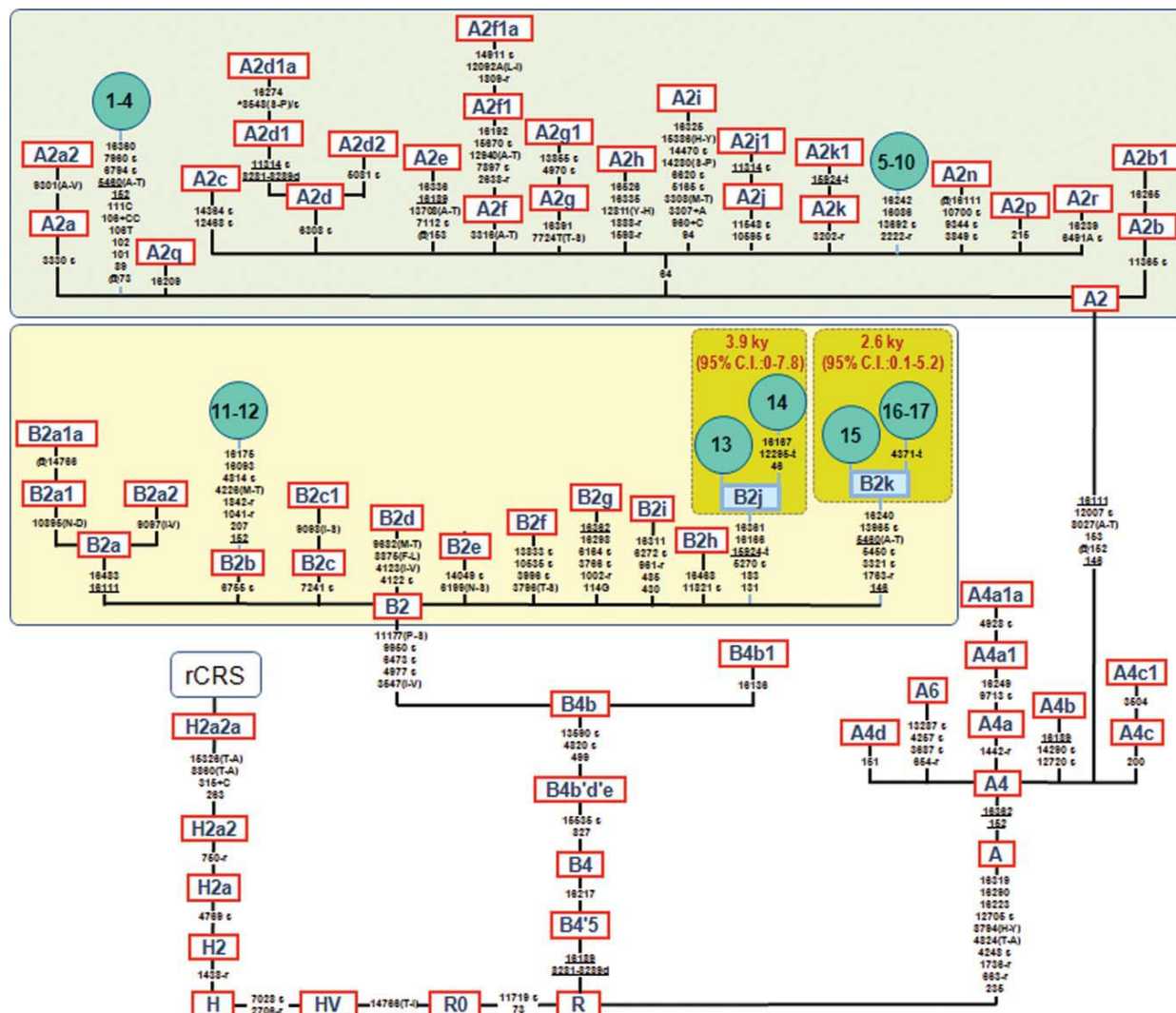


Fig. 2. Skeleton of the phylogeny of hg A2 and B2 complete genomes, as in Phylotree Build 12 (see also Achilli et al. 2008; Tamm et al. 2007), incorporating, on top, the new branches (complete genomes) analyzed in the present study (blue circles). The position of the rCRS is indicated for reading the sequence motif. Mutations are shown on the branches; all positions are transitions unless a base is explicitly indicated as a suffix to signal a transversion or insertion (+). Suffixes indicate: a) transversion (A, G, C, or T), indels (+ or d), gen locus (~ t = tRNA; ~ r = rRNA), synonymous change (s) or non-synonymous changes in brackets. The prefix @ indicates a back mutation while * indicates a position that is located in the overlapping region between the AT6 and AT8 genes. The coalescence age needed to accumulate the variation within B2j and B2k were estimated following Soares et al. (2009). Correspondence between sample codes in the present figure and in Table S1 are as follows: #1-4 = VZ224, VZ208, VZ207, and VZ194; #5-10 = VZ406, VZ402, VZ348, VZ327, VZ316, and VZ308; #11-12 = VZ65 and VZ389; #13 = VEN40; #14 = VEN44; #15 = VEN26; #16-17 = VZ307 and VZ329. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://www.wileyonlinelibrary.com).]

variation located around 16189, which is usually associated with length heteroplasmy, for example 16182C or 16183C, was ignored. Data from America ($n = 6817$), Africa ($n = 13594$), and Europe ($n = 5658$; considering only Portugal, Spain and Italy) were collected from the literature as indicated in Supporting Information Table S2 and (Salas et al., 2009b); most of this data correspond with HVS-I segments of variable length. Full available information was used for phylogeographic descriptive analysis when comparing Venezuelan mtDNAs and those collected from the literature, which also included entire genome data; e.g., from Phylotree.

For the mtDNA complete genomes (Figure 2) analyzed in the present study, estimation of the time of the most

recent common ancestor (TMRCA) and its standard deviation (SD) was done according to Saillard et al. (2000). We used the evolutionary rate estimate for the entire mtDNA molecule reported by Soares et al. (2009), and we also used the calculator tool available in this study.

To estimate the most likely origin of native American lineages in admixed Venezuelans and admixture proportions in the European Venezuelan component as coming from Spain, Portugal, and Italy, we used the admixed model previously described by Mendizabal et al. (2008), based on shared haplotypes between regions. In brief, this model computes the probability of the origin of Venezuelan mtDNAs being in each of the source regions considered by computing $p_{O_s} = \frac{1}{n} \sum_{i=1}^n k_i \frac{B_{is}}{p_{sc}}$ where, n is the num-

ber of Venezuelan sequences with perfect matches (≥ 1) in the whole database; k_i , is the number of times the sequence i is found in Venezuela; p_{is} , is the frequency of the sequence i in each source dataset (e.g., Spain and Italy); and p_{ic} , is the frequency of the sequence i in whole database. Here, we extended the original model of Mendizabal et al. by considering n to be the number of native American Venezuelan sequences that match ($p_{0_s} = P_0$) or have one ($p_{0_s} = P_1$) or two ($p_{0_s} = P_2$) mutational step differences with the sequences contained in the database. Estimates were obtained using a bootstrap procedure aimed to account for sampling errors and different sample sizes in the source populations. A total of 1000 resamples of size 300 were taken as random from the source populations from which we derived the mean and the 95% C.I. of the admixture components. Resamples of different size (ranging from 100 to 600) yielded virtually the same results (data not shown).

RESULTS

mtDNA diversity in Venezuela

The native American mtDNA component of the Venezuelans was 80% of the total sample (considering Caracas and Pueblo Llano together). Most of these mtDNAs were allocated to hg A2 (46%), followed by hg C1 (21%), hg B2 (7%) and hg D1 (6%). As expected, the indigenous component was significantly higher in Pueblo Llano (89%) than in Caracas (65%), whereas Caracas showed a higher European (15%) and African (20%) mtDNA ancestry than Pueblo Llano (6% and 4%, respectively); see Fig. 1. Therefore, patterns of hg frequencies significantly vary between Venezuelan populations; for instance, no hg A2 mtDNAs were observed in the Yanomami while A2 is the predominant clade in other Venezuelan locations (see introduction). In the context of the whole American continent, this demonstrates once more the important role of genetic drift in the settlement and posterior demographic evolution of native American populations. There are many examples in the literature showing dramatic differences in hg frequencies between populations living in neighboring regions (see Fig. 1 in Salas et al., 2009b).

Caracas had higher diversity values than Pueblo Llano, partly due to the presence of a higher African mtDNA ancestry (Supporting Information Table S1).

As indicated by the haplotype (H) and nucleotide diversity (π) indices, as well as the average nucleotide pairwise differences (M), the Venezuelan native American mtDNAs had a relatively low diversity in comparison to the average values observed in North, Meso, and South America (Table 2). Although Pueblo Llano seemed to harbor higher H values than Caracas, the latter had higher values of M and π .

The diversity observed in the African L-mtDNAs from Venezuela was very similar to the values observed in Africa. Pueblo Llano only had eight mtDNAs belonging to L-hgs, and therefore, the diversity estimates might not have been fully representative of the African component of the region.

Phylogeography of urban Venezuelan mtDNAs and admixture proportions

A large database of native American HVS-I profiles (considering the sequence range 16090 to 16365; $n > 7000$) did not contain exact matches for 35 out of the 66 (53%)

different mtDNA haplotypes observed in the admixed Venezuelans (Table 3).

An important percentage of the native American haplotypes were shared between North, Meso, and South America for the sequence range considered; for instance, there were 925 different native American haplotypes in South America, 79 of which were also present in North America and 88 in Meso America (Fig. 3). Some of them were very common in the double continent (Table 3), including Venezuela. A Bayesian-based method was used to infer the proportions of the admixture contributed by the South, Meso, and North American regions. The native American component of native Venezuelans was also included in the model, somehow representing the autochthonous native component of the urban populations analyzed in the present study. According to this model, the main contributor to the urban native mtDNAs most likely came from South American locations, including natives from Venezuela (average contribution: 44%) (Table 4; the components that came from North and Meso America were also important, 29 and 26%, respectively).

The European components of Pueblo Llano and Caracas accounted for ~5 and ~14% of the total mtDNAs, respectively. Only few of these lineages can be tentatively traced to more precise locations within Europe. Thus, exact matches of the two Pueblo Llano J1c profiles with identical HVS-I segment (C16069T T16126C G16145A T16231C C16261T) were observed in several Spanish regions, e.g. (Álvarez et al., 2007) although close phylogenetic relatives are common all around Europe and also in European-American descents.

On the other hand, most of the hg U6 mtDNAs (Olivieri et al., 2006) found in America most likely came from North Africa although with some exceptions. The hg U6 mtDNA observed twice in Caracas, A16163G T16172C A16219G T16311C, undoubtedly belong to the Canary Island branch U6b1; these islands representing the Spanish region that contributed more immigrants to Venezuela. At least 31 exact matches of this mtDNA profile were found in the Canary islands (and in no other African or European locations) and, curiously, also in American locations that received significant numbers of immigrants from these islands in modern times, such as Cuba (Sandoval et al., 2009) and Uruguay (Pagano et al., 2005). The Caracas U6a profile T16172C A16219G C16278T was highly prevalent in North Africa, but it could also have arrived in Venezuela from western Africa (Rando et al. 1998; Brehm et al., 2002) or even, more likely, indirectly through Portugal [where exact matches were also observed (Pereira et al., 2004)], given the important role of this country in providing immigrants to Caracas.

It was also noticeable that exact matches of the Pueblo Llano HVS-I profile T16067C C16292T C16354T were only found in Armenia (Richards and Macaulay, 2000) and in the central region of Saudi Arabia (Abu-Amro et al., 2008), and that exact matches of the Caracas R0a profile T16126C T16189C T16362C were exclusively found in Armenia (Richards and Macaulay, 2000) and in a different Arab population from the Chad Basin (Cerný et al., 2007).

A total of 10% of the mtDNAs of admixed Venezuelans belonged to the typical sub/Saharan L-hgs. As already predicted in previous studies for other American regions (Salas et al., 2004, 2005a) most of the Venezuelan L-hg lineages were most likely attributed to the slave trade

mtDNA IN ADMIXED VENEZUELAN

83

TABLE 3. Native American haplotypes observed in admixed Venezuelans (only HVS-I segment 16090 to 16365) shared with South, Meso, and North America

HVS-I (minus 16000)	Hg	Venezuela (n = 261)	North America (n = 2264)	Meso America (n = 1435)	South America (n = 3340)
092 155 223 290 319	A2	1	—	—	—
111 223 290 319 362	A2	63	239	220	251
111 129 223 290 319 362	A2	26	29	9	44
111 223 290 311 319 362	A2	2	4	6	3
111 223 239 284 290 319 362	A2	2	—	—	—
111 223 290 296 319 362	A2	1	—	—	—
111 129 223 290 293C 319 362	A2	1	—	—	—
111 223 258C 290 319 362	A2	1	—	—	—
111 213 223 290 319 356 362	A2	13	—	1	—
094 111 129 223 290 319 362	A2	1	—	—	—
111 223 242 290 319 362	A2	7	—	—	—
111 223 290 319 359 362	A2	1	—	—	—
111 126 223 256 290 319 362	A2	1	5	5	—
166 189 193+C	B?	1	—	—	—
189 193+C 217	B4b	1	5	1	13
189 193+C 217 221	B4b	1	—	—	—
166 167 189 193+C 217 361	B2j	1	—	—	—
189 193+C 217 234	B4b	1	—	—	—
223 298 327	C1	1	20	3	3
223 292 298 325 327 362	C1	11	—	—	15
129 223 325 327	C1	15	—	—	—
223 298 325 327	C1	8	197	80	221
189 223 298 325 327	C1	1	8	5	13
223 292 298 311 325 327 362	C1	1	—	—	—
223 298 299 325 327	C1	1	3	—	—
223 271 292 298 325 327 362	C1	1	—	—	—
189 193+C 223 278 298 325 327	C1	1	6	—	—
189 193+C 223 298 311 325 327	C1	1	1	—	—
261 298 325 327	C1	1	—	—	—
142 179 223 295 325 362	D1f	8	1	—	20
223 289 325 362	D1	1	—	—	—
092 142 223 311 325 362	D1f	3	—	—	—
223 325 362	D1	1	46	28	91
189 217 240	B2k	3	—	—	—
166 189 217 361	B2j	1	—	—	—
166 167 189 217 361	B2j	1	—	—	—
223 298 325 327 362	C1	1	4	1	13
111 223 242 290 319 355 362	A2	1	—	—	—
092 111 223 290 319	A2	1	—	—	1
099 111 223 290 319 362	A2	1	—	—	—
111 126 223 290 319 362	A2	1	—	4	—
111 129 212 223 234 258 290 319 362	A2	1	—	—	—
111 129 223 256 290 319 362	A2	1	—	1	—
111 155 223 290 319	A2	1	—	—	—
111 172 223 290 319 362	A2	1	1	14	9
111 189 223 290 319 362	A2	3	5	19	30
111 213 223 290 319 355 362	A2	2	—	—	—
111 213 223 290 319 362	A2	7	1	13	3
111 223 256 290 319 362	A2	4	16	1	—
111 223 290 319 360 362	A2l	5	1	22	2
111 189 223 242 290 319 355 362	A2m	1	—	—	—
093 175 189 217	B2b1	4	—	—	—
148 189 217	B4b	1	2	—	—
178 189 217	B4b	2	—	—	15
189 217	B4b	5	78	87	241
189	?	1	—	—	3
189 223 325 327	C1	3	—	—	1
129 189	C1	4	—	—	—
129 189 193 223 325 327	C1	1	—	—	—
129 189 223 325 327	C1	8	—	—	—
129 223 325 327 362	C1	6	—	—	—
223 325 327	C1	1	4	3	29
129 189 223 325 327 362	C1	1	—	—	—
189 223 325 362	D1	1	3	8	15
092 142 223 325 362	D1f	1	—	—	—
187 203 223 241 245 301 319 342 362	D4h3a	6	—	—	—

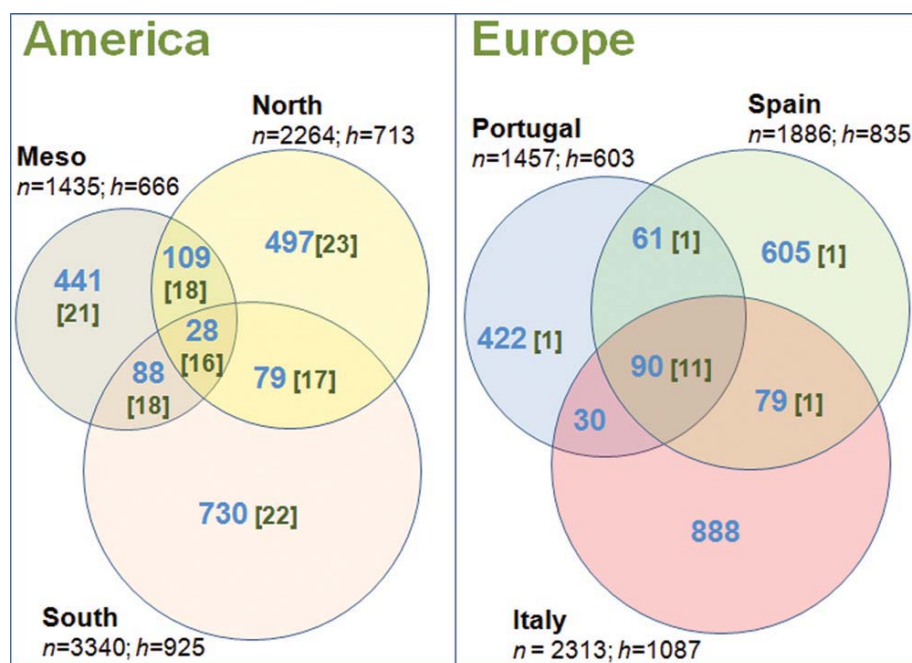


Fig. 3. Haplotypes shared between the European mtDNA component of Venezuela and the main European contributors (Spain, Portugal, and Italy) and the native American component of Venezuela *versus* the native American component of North, Meso, and South America. The numbers in blue indicate the number of different haplotypes; numbers in square brackets are the numbers of different Venezuelan haplotypes that were present in the corresponding populations; n = sample size of each population; h = number of different haplotypes in each population. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 4. Bayesian-based estimated contribution of different American regions to the Native American component of admixed Venezuelans

Continental region	P_0	95% C.I. (P_0)	P_1	95% C.I. (P_1)	P_2	95% C.I. (P_2)
America						
North America ($n = 2260$)	0.2863	0.2834–0.2893	0.2680	0.2651–0.2709	0.2847	0.2817–0.2876
Meso America ($n = 1475$)	0.2830	0.2800–0.2859	0.3132	0.3101–0.3162	0.3344	0.3314–0.3375
South America ($n = 3082$)	0.4307	0.4275–0.4339	0.4189	0.4156–0.4221	0.3809	0.3777–0.3840
Europe						
Spain ($n = 1886$)	0.3847	0.3815–0.3879	0.4022	0.3990–0.4054	0.3715	0.3684–0.3747
Portugal ($n = 1457$)	0.3452	0.3421–0.3483	0.3091	0.3061–0.3121	0.3291	0.3261–0.3322
Italy ($n = 2313$)	0.2701	0.2672–0.2730	0.2887	0.2858–0.2917	0.2993	0.2964–0.3023

See the text for more information on P_0 , P_1 , and P_2 .

coming from the West African coast. For instance, haplotype C16114G T16126C C16187T T16189C C16223T C16264T C16270T G16274A C16278T A16293G T16311C (hg L1b1) was only observed in Senegal (Graven et al., 1995; Rando et al., 1998) and Cape Verde (Brehm et al., 2002). Haplotype T16129C C16148T T16172C C16187T C16188G T16189C C16192T C16223T A16230G C16291T T16311C A16317G C16320T belonging to L0a1 appeared five times in Africa, all of them from the Kota from Gabon (Quintana-Murci et al., 2008), and twice in Caracas. The three L1c mtDNAs were also of Western or West/Central African origin (Quintana-Murci et al., 2008). The L3b lineage T16124C C16223T C16278T T16362C had a predominant distribution in Western or West/Central Africa. The lineage T16172C C16189T C16223T C16320T belonging to L3e2b was only found in a sample from two neighboring African/American popu-

lations, three times in Brazil (Bortolini et al., 1997) and nine times in the Central American Garífunas (Salas et al., 2005b). The mtDNA hg L3h profile G16129A C16223T C16256A T16311C T16362C was only observed in one sample from Cabo Verde (Brehm et al., 2002) and one 'Afro Brazilian' sample (Carvalho et al., 2008).

When using an admixture model to infer the estimated proportions of the main European contributors to Venezuela, we observed a good correspondence between the official census (Table 1) and the mtDNA genetic estimates (Table 4). Thus, according to the census, Spain, Portugal, and Italy, contributed 91% of the total European immigrants to Venezuela in the following proportions: 39, 27, and 25% (of the total Europeans). The admixture model indicated a contribution of ~38.4, ~34.5 and ~27.0% from Spain, Portugal, and Italy, respectively. Therefore, estimates derived from mtDNA

data surprisingly fit well with the ones inferred from the census; the slight differences could be due to sex bias, which seems to have more effects in other locations such as in Argentina (Catelli et al., 2011) and in other American locations (Salas et al., 2008b, 2009b).

New branches of the native American hg A2 and B2

Five new branches of the native American phylogeny were revealed by sequencing 17 entire mtDNA genomes (Fig. 2). Three of them belonged to hg B2, whereas two were new branches of hg A2.

One of the A2 branches shows 12 variants with respect to the root and has a very characteristic diagnostic HVS-II motif (Fig. 2): T89C G101A A102G G106T 106+TT A111C T152C (plus a reversion at position 73). We have observed four entire genomes that displayed the exact same sequence. Interestingly, no matches were found for this sequence in a large database of more than 1600 HVS-II American sequences (since it is in the HVS-II that this lineage has a clear diagnostic motif), but it appeared six times in Pueblo Llano.

Also, strikingly is that the six entire genomes analyzed that define the other A2 branch, showed the same identical entire genome sequence (Fig. 2). When searching its specific HVS-I motif (transitions at 16086 and 16242) in a database of >14,600 American sequences (for which the HVS-I data is available), this branch is further supported by eight unrelated individuals all observed in Pueblo Llano, one sample from Caracas (Lander et al., 2008), one African-Brazilian from Tamauari in a northern Brazilian location (Maranhao state) (Carvalho et al., 2008), and one from Nahua in Xochimilco (Mexico; (Sandoval et al., 2009)).

Two entire genomes belong to a new branch of hg B2b and were also sampled in Pueblo Llano. Again, these genomes were identical, probably as a consequence of the same evolutionary forces that lead to the null genetic variation observed within two novel A2 branches. The HVS-I motif was again highly prevalent in Pueblo Llano (four individuals), but it also appeared in a 'Mestizo' sample from Colombia (Salas et al., 2008a), in another one from Costa Rica (Morera, 2002) and in two from the Genographic project (<https://genographic.nationalgeographic.com/genographic/index.html>).

Haplogroup B2j is characterized by the mutations T131C A183G C5270T A15824G A16166G A16183C G16361A T16519C. A search of its diagnostic HVS-I motif (A16166G G16361A) in an extensive American HVS-I database only yielded one additional mtDNA outside Venezuela; reported by the Genographic project. Its coalescence age was estimated as 3.9 ky (95% C.I: 0–7.8); therefore, it is a recent clade within the phylogeny of hg B2 (Achilli et al., 2008; Soares et al., 2009).

An additional young branch of B2, labeled here as B2k, was suggested by other two complete genomes; its diagnostic position A16240G in HVS-I identified another mtDNA outside Venezuela in the Apache population reported by Monson et al. (2002), and it seems to have a similar (overlapping) coalescence age as B2j, namely, 2.6 ky (95% C.I: 0.1–5.2).

DISCUSSION

According to the last official census carried out in Venezuela (2001; INI), there are about 530,000 indigenous people living in Venezuela, representing 2.3% of the population. About 33% of them live in indigenous communities grouped into 28 ethnic groups. The Guahiro

or Wayúu constitute the most important community (>50% of the indigenous communities in Venezuela). Although the native Venezuelans communities represent a small part of the Venezuelan population, this study shows that the native American mtDNA component is by far the most prevalent in present day urban Venezuelans (80%), whereas it is much more frequent in Pueblo Llano (~90%) than in Caracas (~65%). The different proportions observed in the two urban populations fit well with their documented demographic history. Thus, the first inhabitants of Pueblo Llano were the Chinoes and the first contact with the Spaniards was in 1559; until very recently, the region has been relatively isolated from other more populated areas of Venezuela (Picón-Parra, 1988). Caracas, however, has been receiving European immigrants almost continuously since colonial times. For instance, after the Second World War, more than 210,000 Italians arrived in Caracas, mainly from Sicily, Campania, Abruzzo and Apulia (INE). Massive waves of Spaniards and Portuguese also arrived in Venezuela, since Caracas was one of the main gateways for European immigrants (Table 1). The admixed proportions obtained in this study add to the complexity already observed in the literature in other Latin-American communities (Table S3); some of them with a predominant African ancestry, e.g., Garífunas from (Salas et al., 2005b); whereas others have a more prevalent native American component e.g., Mexicans (Green et al., 2000).

Although urban Venezuelans still preserve the legacy of their indigenous inhabitants, other American locations, especially neighboring countries (e.g., Colombia) and the Caribbean are also likely to have contributed to their native American component in recent times, as indicated by the admixture analyses of the present study (Table 4). Immigration from countries with an important native American component, such as Colombia, Ecuador, Chile, Dominican Republic, Cuba, Guyana, for example, is well documented in the official census [<http://www.ine.gov.ve/demografica/>; (Peyser and Chakiel, 1999)]. Compatible with the nonautochthonous nature of many native American urban profiles, the vast majority of the native American mtDNAs in urban Venezuela were not observed in the three indigenous populations analyzed in the region to date (Merriwether et al., 2000; Williams et al., 2002; Vona et al., 2005). The approach carried out in this study to estimate admixture proportions was based on haplotype sharing since other methods based on hg frequencies are inadequate given the low resolution provided by native American haplogroups (see Materials and Methods). Exclusively based on the analysis of the HVS-I segment it is not possible to objectively evaluate which proportion or to what extent the haplotypes are shared due to common shared ancestry from the proportion that is shared with neighboring countries due to recent migration. The inclusion of native Venezuelan lineages in the admixture model was intended to contribute to this differentiation by providing autochthonous lineages to the model. However, these estimates should be taken with caution awaiting further studies based (ideally) on complete mtDNA genome data and other molecular markers.

The autochthonous nature of some native American lineages observed in urban Venezuelans was more clearly revealed through the selective analysis of entire mtDNA genomes. Entire genome sequencing of 17 mtDNAs allowed the discovery of five new branches of the native American phylogeny (Fig. 2). The two A2

branches (and to a lesser extent the new reported sub-branch of B2b) show the imprint of severe genetic drift and/or consanguinity in Pueblo Llano. A thorough search of the HVS-I motifs of these lineages in a large native American database indicated that they hgs are most likely autochthonous in Venezuela (perhaps from the Pueblo Llano area), but further studies are needed in order to determine more about the variation at the level of entire genomes in other American locations. The immediate ancestors of these lineages could have been evolved in other American regions; but the search for their control region motifs in a large continental database indicated the existence of close phylogenetically related sequences only in Colombia, Brazil, Costa Rica, and Mexico. On the other hand, hgs B2j and B2k evolved locally in Venezuela, and could be dated to about 3.9 (95% C.I. 0–7.8) and 2.6 (95% C.I. 0.1–5.2) kya.

The phylogenetic patterns observed through the analysis of entire genomes, coupled with the reduced values of the genetic diversity of mtDNA observed in Pueblo Llano (Table 2) compared with Caracas, represent a clear imprint of severe recent genetic bottlenecks in this region. High consanguinity, as documented in the region, also contributed to the reduced variability observed at the mtDNA level. The latter would explain the high prevalence of Mendelian and complex diseases suffered by the Pueblo Llano.

The majority of the African L-hgs lineages observed in Venezuela most likely came from West-Central Africa, as previously reported for other American locations and populations (Pollak-Eltz, 1972; Thornton, 1998; Mendizabal et al., 2008; Salas et al., 2008a, 2009b; Sandoval et al., 2009).

We were able to determine the most likely geographical origins of some of the European mtDNAs. Thus, for instance, the presence of two U6b1 haplotypes clearly testified for the arrival of immigrants from the Canary Islands.

Variations of the mtDNA genome have shown that urban Venezuela represents a melting pot of different ancestries, where African and European strata, as well as the native American component of other neighboring countries, have become superimposed on its original indigenous background. The analysis of two different admixed populations from this country, Caracas and Pueblo Llano, revealed the existence of a substantial stratification within urban Venezuela. Population stratification is a common confounding effect in the population-based studies typically employed today to analyze genetic predispositions for complex and common diseases. Although mitochondrial variation alone is not enough to control or monitor population stratification (Mosquera-Miguel et al., 2008; Salas et al., 2009a), it could be useful for highlighting the existence of genetic heterogeneity in a given population.

ACKNOWLEDGMENTS

The authors like to thank the donors for their participation in the present project. There were no conflicts of interest in this study. The complete genomes analyzed in the present study have been submitted to GenBank under the accession numbers JF431049–JF431065.

LITERATURE CITED

- Abu-Amro KK, Larruga JM, Cabrera VM, González AM. 2008. Mitochondrial DNA structure in the Arabian Peninsula. *BMC Evol Biol* 8:45.
- Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, Woodward SR, Salas A, Torroni A, Bandelt H-J. 2008. The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. *PLoS ONE* 3:e1764.
- Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A. 2007. Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. *Forensic Sci Int Genet* 1:44–55.
- Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, Cuscó I, Lareu MV, García O, Pérez-Jurado L, Carracedo Á, Salas A. 2009. New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. *PLoS One* 4:e5112.
- Álvarez JC, Johnson DE, Lorente JA, Martínez-espín E, Martínez-González LJ, Allard MW, Wilson MR, Budowle B. 2007. Characterization of human control region sequences for Spanish individuals in a forensic mtDNA data set. *Legal Med* 9:293–304.
- Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23:147.
- Bandelt H-J, Kivisild T, Parik J, Villems R, Bravi CM, Yao Y-G, Brandstätter A, Parson W. 2006. Lab-specific mutation processes. In: Bandelt H-J, Richards M, Macaulay V, editors. *Human mitochondrial DNA and the evolution of Homo sapiens*. Berlin: Springer-Verlag.
- Bortolini MC, Zago MA, Salzano FM, Silva-Junior WA, Bonatto SL, da Silva MC, Weimer TA. 1997. Evolutionary and anthropological implications of mitochondrial DNA variation in African Brazilian populations. *Hum Biol* 69:141–159.
- Brehm A, Pereira L, Bandelt H-J, Prata MJ, Amorim A. 2002. Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade. *Ann Hum Genet* 66:49–60.
- Brisighelli F, Capelli C, Álvarez-Iglesias V, Onofri V, Paoli G, Tofanelli S, Carracedo Á, Pascali VL, Salas A. 2009. The Etruscan timeline: a recent Anatolian connection. *Eur J Hum Genet* 17:693–696.
- Carvalho BM, Bortolini MC, S.E. BdS, Ribeiro-dos-Santos ÁKC. 2008. Mitochondrial DNA mapping of social-biological interactions in Brazilian Amazonian African-descendant populations. *Genet Mol Biol* 31:12–22.
- Catelli ML, Álvarez-Iglesias V, Gómez-Carballa A, Mosquera-Miguel A, Romanini C, Borosky A, Amigo J, Carracedo Á, Vullo C, Salas A. 2011. The impact of modern migrations on present-day multi-ethnic Argentina as recorded on the mitochondrial DNA genome. *BMC Genet* 21:77.
- Cerezo M, Cerný V, Carracedo Á, Salas A. 2011. New insights into the Lake Chad Basin population structure revealed by high-throughput genotyping of mitochondrial DNA coding SNPs. *PLoS ONE* 6:e18682.
- Cerný V, Salas A, Hájek M, Žaloudková M, Brdicka R. 2007. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. *Ann Hum Genet* 71:433–452.
- Dillehay T. 2000. *The settlement of the Americas*. Oxford: Marston Book Services Limited.
- Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA Jr., Zago MA, Ribeiro-dos-Santos AK, et al. 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet* 82:583–592.
- Graven L, Passarino G, Semino O, Boursot P, Santachiara-Benerecetti S, Langaney A, Excoffier L. 1995. Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol Biol Evol* 12:334–345.

- Green LD, Derr JN, Knight A. 2000. mtDNA affinities of the peoples of North-Central Mexico. *Am J Hum Genet* 66:989–998.
- Lander N, Rojas MG, Chiurillo MA, Ramirez JL. 2008. Haplotype diversity in human mitochondrial DNA hypervariable regions I–III in the city of Caracas (Venezuela). *Forensic Sci Int Genet* 2:e61–64.
- Lewis MP. 2009. *Ethnologue. Languages of the world*. Dallas, TX: SIR International.
- Martínez H, Rodríguez-Laralde A, Izaguirre MH, De Guerra DC. 2007. Admixture estimates for Caracas, Venezuela, based on autosomal, Y-Chromosome, and mtDNA markers. *Hum Biol* 79:201–213.
- Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martínez-Fuentes A, Comas D. 2008. Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. *BMC Evol Biol* 8:213.
- Merriwether DA, Kemp BM, Crews DE, Neel JV. 2000. Gene flow and genetic variation in the Yanomama as revealed by mitochondrial DNA. In: Renfrew C, editor. *America past, America present: genes and languages in the Americas and beyond*. Cambridge: McDonald Institute for Archaeological Research. p 89–124.
- Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B. 2002. The mtDNA Population Database: an integrated software and database resource for forensic comparison. *Forensic Sci Commun* 4:2.
- Morera B. 2002. Análisis del polimorfismo del ADNmt en la población general de Costa Rica: un asunto pendiente. *Revista Latinoamericana de Derecho Médico y Medicina Legal* 7:21–34.
- Morón G. 1994. *Breve historia contemporánea de Venezuela*. Mexico: Fondo de Cultura Económica.
- Mosquera-Miguel A, Álvarez-Iglesias V, Vega A, Milne R, Cabrera de León A, Benítez J, Carracedo A, Salas A. 2008. Is mitochondrial DNA variation associated with sporadic breast cancer risk? *Cancer Res* 68:623–625.
- Nei N. 1987. *Molecular evolutionary genetics*. New York: Columbia University Press.
- Oschenius C, Gruhn R. (eds.) (1979). *Taima-Taima: A late Pleistocene kill-site in northernmost South America*, CIPICS/South American Quaternary Documentation Program, Coro, Venezuela.
- Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM et al. 2006. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science* 314:1767–1770.
- Pagano S, Sans M, Pimenoff V, Cantera AM, Álvarez JC, Lorente JA, Peco JM, Mones P, Sajantila A. 2005. Assessment of HV1 and HV2 mtDNA variation for forensic purposes in an Uruguayan population sample. *J Forensic Sci* 50:1239–1242.
- Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Kashani BH, Ritchie KH, Scozzari R, Kong Q-P, et al. 2009. Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr Biol* 19:1–8.
- Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Kashani BH, Carossa V, Ekins JE, Gomez-Carballea A, Huber G, et al. 2010. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. *Genome Res* 20:1174–1179.
- Pereira L, Cunha C, Amorim A. 2004. Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. *Int J Legal Med* 118:132–136.
- Peyser A, Chakiel J. 1999. La identificación de poblaciones indígenas en los censos de América Latina. Santiago de Chile: CEPAL/CELADE.
- Picón-Parra R. 1988. *Fundadores, Primeros Moradores y Familias Coloniales de Mérida*. Caracas, Distrito Federal, Venezuela.
- Pollak-Eltz A. 1972. *Procedencia de los esclavos negros traídos a Venezuela*. Caracas Universidad Católica Andrés Bello, Instituto de investigaciones Históricas.
- Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mougouma-Daouda P, Comas D, Tzur S, et al. 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci USA* 105:1596–1601.
- Rando JC, Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM, Bandelt H-J. 1998. Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, Near-Eastern, and sub-Saharan populations. *Ann Hum Genet* 62:531–550.
- Richards M, Macaulay V. 2000. Genetic data and the colonization of Europe: genealogies and founders. In: Renfrew C, Boyle K, editors. *Archaeogenetics: DNA and the population prehistory of Europe*. Cambridge: McDonald Institute for Archaeological Research. p 139–151.
- Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.
- Saillard J, Forster P, Lynnerup N, Bandelt H-J, Norby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* 67:718–726.
- Salas A, Acosta A, Álvarez-Iglesias V, Cerezo M, Phillips C, Lareu MV, Carracedo A. 2008a. The mtDNA ancestry of admixed Colombian populations. *Am J Hum Biol* 20:584–591.
- Salas A, Carracedo A, Richards M, Macaulay V. 2005a. Charting the ancestry of African Americans. *Am J Hum Genet* 77:676–680.
- Salas A, Fachal L, Marcos-Alonso S, Vega A, Martínón-Torres F, ESIGEM G. 2009a. Investigating the role of mitochondrial haplogroups in genetic predisposition to meningococcal disease. *PLoS One* 4(12):e8347.
- Salas A, Jaime JC, Álvarez-Iglesias V, Carracedo A. 2008b. Gender bias in the multi-ethnic genetic composition of Central Argentina. *J Hum Genet* 53:662–674.
- Salas A, Lovo-Gómez J, Álvarez-Iglesias V, Cerezo M, Lareu MV, Macaulay V, Richards MB, Carracedo A. 2009b. Mitochondrial echoes of first settlement and genetic continuity in El Salvador. *PLoS One* 4:e6882.
- Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A. 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am J Hum Genet* 74:454–465.
- Salas A, Richards M, Lareu MV, Sobrino B, Silva S, Matamoros M, Macaulay V, Carracedo A. 2005b. Shipwrecks and founder effects: divergent demographic histories reflected in Caribbean mtDNA. *Am J Phys Anthropol* 128:855–860.
- Sandoval K, Buentello-Malo L, Peñaloza-Espinosa R, Avelino H, Salas A, Calafell F, Comas D. 2009. Linguistic and maternal genetic diversity are not correlated in Native Mexicans. *Hum Genet* 126:521–531.
- Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* 84:740–759.
- Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.
- Tajima F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. *Mol Biol Evol* 10: 77–688.
- Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martínez-Labarga C, Khushnudinova EK et al. 2007. Beringian standstill and spread of Native American founders. *PLoS ONE* 2(9):e829.
- Thornton J. 1998. *Africa and Africans in the making of the Atlantic world, 1400–1800*. Cambridge: Cambridge University Press.
- Van Oven M, Kayser M. 2008. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat*. In press.
- Vona G, Falchi A, Moral P, Calo CM, Varesi L. 2005. Mitochondrial sequence variation in the Guahibo Amerindian population from Venezuela. *Am J Phys Anthropol* 127:361–369.
- Williams SR, Chagnon NA, Spielman RS. 2002. Nuclear and mitochondrial genetic variation in the Yanomamo: a test case for ancient DNA studies of prehistoric populations. *Am J Phys Anthropol* 117:246–259.

Bloque 1.

V.4. PIMA: A population indicative multiplex for the Americas

Christopher Phillips, Liliana Porras-Hurtado, Ana Freire-Aradas, Manuel Fondevila, Carla Santos, Antonio Salas, Julieta Henao, Carlos Isaza, Leonardo Beltrán, Vivian Nogueira Silbiger, Adriana Castillo, Adriana Ibarra, Fabien Moreno Chavez, Jens Sochtig, Yarimar Ruiz, Carolina Gontijo, Silviene F de Oliveira, Guillermo Barreto, Fernando Rondon, William Zabala, Lisbeth Borjas, Ángel Carracedo, Maria Victoria Lareu .

(Manuscript in preparation)

PIMA: Population Informative Multiplex for the Americas

Christopher Phillips^{a*§}, Liliana Porras-Hurtado^{ab*}, Ana Freire-Aradas^a, Manuel Fondevila^a, Carla Santos^a, Antonio Salas^a, Julieta Henao^b, Carlos Isaza^b, Leonardo Beltrán^b, Vivian Nogueira Silbiger^c, Adriana Castillo^d, Adriana Ibarra^e, Fabien Moreno Chavez^f, Jens Sochtig^a, Yarimar Ruiz^a, Carolina Gontijo^{a,i}, Silviene F de Oliveiraⁱ, Guillermo Barreto^g, Fernando Rondon^{dg}, William Zabala^h, Lisbeth Borjas^h, Ángel Carracedo^a, Maria Victoria Lareu^a

a Forensic Genetics Unit, University of Santiago de Compostela, Spain

b Medical Genetic Laboratory, Human Molecular Genetic Research Group, Technology University of Pereira, Colombia

c Faculty of Pharmaceutical Sciences, University of Sao Paulo, Federal University of Rio Grande do Norte (Brazil).

d Medical Genetic Laboratory, Industrial University of Santander (UIS) (Colombia)

e Medical Genetic Laboratory, University of Antioquia (Colombia)

f Laboratory of Chile

g Human Molecular Genetic Research Group, University of Valle, Colombia

h Molecular Genetic Laboratory, Medical Genetics Unit, University of Zulia, Venezuela

i Laboratório de Genética, Departamento de Genética e Morfologia, Instituto de Ciências Biológicas, Universidade de Brasília (Brazil)

**These authors contributed equally to this work*

§ Corresponding author. Tel.: +34 981582327; fax: +34 981580336.

E-mail address: c.phillips@mac.com (C. Phillips)

Abstract:

With the goal of building an ancestry-informative autosomal SNP set as a small-scale, flexible multiplex for the analysis of American populations' variability, we have chosen loci showing extreme allele frequency differences when comparing Native Americans to Africans, Europeans, and East Asians. The resulting single multiplex: PIMA (Population Informative Multiplex for the Americas), comprises a set 27 autosomic SNPs plus a gender-specific marker that complements the existing 34plex SNP ancestry test (Phillips et al., 2007), to provide a powerful and simple tool for the analysis of American populations, including those showing a range of complex admixture histories. We have analyzed geographically and culturally diverse populations from North, South, Central America, and the Caribbean from different databases and compared them to our study set. In the latter, we were able to compare the patterns of admixture detected by our compact marker set to those obtained by another panel with a much larger number of markers and to gauge the efficiency of the admixture analyses in each case. Finally, we performed a comparison between the biogeographical ancestry inferred by the presented AIM-SNPs to lineage markers (mtDNA and Y-chromosome).

Introduction:

America is the continent representing the most complex ancestral genetic background nowadays. Demographic movements along history have determined their high heterogeneity and level of admixture. The peopling of this continent occurred from the extreme northeast of Siberia (Beringia) as the source of the first founders, with succeeding waves of migration from more southerly parts of East Asia, over the approximated period of 14,000-18,000 years ago (Volodko, 2008). Later in the XV century, the arrival of European colonizers, as well as the African slave trade, lead to the introduction and subsequent expansion of these ancestral components into the Americas (Malamud, 1993). Such history of migration, admixture, and founder effect resulted in the highly admixed American populations observed nowadays. Before World War II, and more markedly after it, new migratory waves from Eurasia (namely Europe, Middle East, and East Asia) represented an important influx of people - and hence genetic diversity - into the continent. (Salzano, 2002).

The available information from the analysis of Y chromosome and mtDNA in American populations provides a different point of view to their evolutive past in comparison to that provided by autosomal markers: Lineage markers draw the phylogeny of populations (Shields et al., 1993; Underhill et al., 1996; Achilli et al., 2008; Phillips et al., 2008), while autosomal Ancestral Informative Markers (AIMs) allow the study of the modern demographic structure of a given population (Galanter et al., 2012) and provide measurement of admixed ancestry at the individual level. The importance of simultaneous analysis of both lineage and autosomal markers has been reported in previous studies in South American populations (Bedoya 2009; Salas, 2008; Carvajal- Carmona, 2000).

Inference of the biogeographical ancestry (BGA) plays an important role in case-control studies correcting substructure effects in order to avoid false positive results (Pritchard-Donnelly, 2001). In the population genetics field, AIMs help to estimate ancestry proportions in admixed populations, assess their structure and ascertain their recent demographic history. Moreover, AIMs are of great interest in forensic genetics in view of criminal investigation frameworks where no suspects are available. Regarding these complex scenarios, an ancestry profile obtained from a sample could help to reduce the scope of the search and thus, prediction of the geographical origin could indirectly infer phenotypic traits of the perpetrator (Frudakis, 2008).

Several studies reporting AIMs sets able to detect Native-American component (varying in marker number, molecular category and genotyping strategy), have been published to date (Mao et al 2007; Halder et al 2008; Pereira et al 2012, Galanter, 2012). Furthermore, even though some of these panels present interesting approaches in the study of admixed three--component populations (Native American, European, and African) (Galanter et al., 2012), they could yield biased estimates if the population under study presents a recent history of migration from groups not included in the scope of the markers - Asians, for instance. Actually, choosing suitable markers for an American informative panel, especially those able to differentiate between Americans and East Asians, is a difficult task, due to their recent isolation. An example is an indels panel suitable for detecting African, European, East Asian and American components that, despite the good separation between the four ancestral population groups, failed

to correctly classify Siberian Yakuts, which were assigned as Americans (Pereira, 2012).

In the present study, we have developed a *Population Informative Multiplex for the Americas* (PIMA), comprising 27 AIM-SNPs and a gender-specific marker (amelogenin) with genomic positions complimenting those occupied by the markers previously established in a 34-plex assay that focused on the differentiation of Sub-Saharan Africans, Europeans, and East Asians (Phillips et al., 2007). Our goal is bring together the most informative SNP markers for differentiating Native Americans from the other principal continental populations, namely Africa, Europe, and East Asia. Simultaneous analysis of both PIMA and 34plex in populations from North, Central, South America, and the Caribbean were performed and provided broader ranges of admixture estimates. Compared to another panel designed to differentiate Americans, the combined analysis of PIMA and 34plex yielded comparable classification power analyzing a small AIM set. Further, ancestry information from mtDNA and Y-chromosome markers was compared to that provided by the SNP set mentioned above for several populations.

Materials and Methods:

Population samples

Samples were collected with informed consent in all cases. A total of 829 DNA samples were used in this study. Tribal populations from Colombia: Awa (n=32), Coyaima (n=23), Embera (n=7), Pastos (n=33), Pijao (n=32), Mulalo (n=34); Guatemala: Q'eqchi' (n=15) and Venezuela: Wayu (n=28). Admixed populations from Brazil: São Paulo (n=162), Riacho de Sacutiaba (n=29), Kalunga (n=69); Colombia: North West Colombia (n=208), Bucaramanga (n=19); Chile: North Chile (n=25), South Chile (n=30); Guatemala: Guatemala City (n=20) and Venezuela: Maracaibo (n=29), Caracas (n=34). In addition, four admixed populations from Puerto Rico (PUR), Mexico (MXL), Colombia (CLM) and Afro-Americans (ASW) recently released from 1000 Genomes data were included into the study group. A total of 851 samples, which include most of the HapMap database (www.hapmap.org), from three continental population groups (Africa, Europe, and East Asia) were used to construct our reference population set. For the Native American reference group 57 samples

were chosen from the CEPH-HGDP that matched those previously studied using 377 autosomal STRs by Rosenberg et al. (Rosenberg, 2001). The same reference samples from HapMap were used in a comparative analysis aiming to test the predictive power for three and four population differentiation regarding PIMA/34Plex and a recently published panel (Galanter et al., 2012). The only exception was the Native American reference population, which was not available for this data base and was hence replaced by Natives from Venezuela (Amazonia, n=57) that showed a very high Native component compared to GWAs data (Galanter, 2012) for the large panel, and the Americans from CEPH-HGDP (n=57) for the PIMA/34Plex analysis. Some of the study populations were also included in this comparative analysis signaled in Figure 1. Furthermore, reference samples from five continental group chosen from the CEPH-HGDP panel were used in an additional analysis comparing our data set to a previously published Indels panel (Pereira, 2012). Geographic location from all study and reference population from America are represented in **Figure 1**.

AIM-SNP selection

Twenty-seven of the PIMA markers were selected from the largest collection of SNPs (650,000) and global populations (53) characterized to date: the CEPH Human Genetic Diversity Panel (CEPH-HGDP) study by Stanford University (Li, 2008). The Stanford SNP genotypes are publicly available from The CEPH Foundation (Cann, 2002) but we made use of our own online SNP allele frequency browser: SPSmart (Amigo, 2008; Amigo et al., 2011), which allows the simultaneous comparison of the Stanford data with genotypes from HapMap, Perlegen and dbSNP. This generated a candidate pool of over 600 SNPs from which we selected those providing optimum genomic distribution and relatively evenly spaced from a previous 34 SNP selection (minimum 10Mb inter-marker distances). Emphasis was made to select SNPs that could adequately differentiate East Asian and American populations. A single SNP without CEPH genotype data: rs6993205 was selected from HapMap data as this marker showed strong differentiation between African and non-African populations. The resulting PIMA multiplex also included a single X-chromosome SNP and a gender-specific marker for amelogenin. PIMA and the 34-plex assays provided a small set of SNPs representing a near optimum combination of markers at this level of multiplexing for the differentiation of the four major population groups.

Genotyping protocol

PIMA multiplex reaction was optimized using QIAGEN® Multiplex PCR kit, adding 2 µl of PCR Master Mix, 2 µl of primer mix (0.125-2,5 µM), and 1-20 ng of DNA in a final volume of 5 µl. Amplification conditions were: initial denaturation at 95 °C for 15 min; 30 cycles at 94 °C for 30 s, 60 °C for 1 min and 72 °C for 30 s; and a final extension at 72 °C for 10 min. Excess primers and dNTPs were removed by adding 1 µl of ExoSAP-IT (USB® Corporation) to 2.5 µl of the PCR product and incubated at 37 °C for 45 min followed by 85 °C for 15 min to inactivate the enzyme. Single base extension reactions were performed in a final volume of 3 µl containing 1.25 µl of SNaPshot™ reaction mix (AB), 0.75 µl of SBE primer mix and 1 µl of purified PCR product. The SBE primer mix was diluted in 160 mM ammonium sulphate to avoid non-specific hybridizations amongst the primers. The SBE reaction was performed in an AB 9700 thermal cycler with the following cycle program: 30 cycles of 96 °C for 10 s, 55 °C for 5 s and 60 °C for 30 s. Excess nucleotides were removed by addition of 1 µl SAP (1 U/ml Shrimp Alkaline Phosphatase) to the total volume of the extension products and incubation at 37 °C for 80 min and 85 °C for 15 min. A combination of 1 µl of sample, 9.5 µl LIZ 120 size standard plus HiDi formamide at a ratio of 1:33.3 (AB) was analyzed by capillary electrophoresis using an AB 3130 Genetic Analyzer with POP4 or POP6 polymer and analyzed with GeneMapper v4.0. Predefined size windows for each allele were determined from prior analysis of a minimum of 20 samples. In addition the control region of mtDNA was sequenced in some of the samples, following previously developed protocols (Alvarez-Iglesias V, 2009). Then, according to the main haplogroups identified, specific continental mtSNP-plexes were used (Alvarez-Iglesias V et al., 2007), also was employed mtDNA data from Venezuelan published recently (Gomez-Carballa et al., 2012). Finally, the Y-chromosome analyses comprised adaptations from a Y-SNP plex previously published (Brion et al., 1994), and also was used data published recently (Carvajal-Carmona, 2000).

Statistical analysis

Allele frequencies, Hardy-Weinberg equilibrium (HWE) and pairwise F_{st} comparisons were estimated using Arlequin 3.5.1.2. (Excoffier, 2005) Ancestry inferences were performed using Structure v2.3.3 (Pritchard et al., 2000) with a burn-

in length of 200.000 followed by 200.000 MCM and considering an admixture model with correlated allele frequencies and POPFLAG. Three independent runs were analyzed for each testing K value, ranging from K=2 to K=7 (number of assumed clusters). CLUMPP v1.1.2 (Jakobsson-Rosenberg, 2007) was used to obtain the average permuted individual and population Q-matrices throughout the three replicates for each K value. Those matrices were used as input to distruct v1.1 (Rosenberg, 2004) to obtain the reported bar plots. Principal Components Analysis (PCA) was executed using R software v2.13.1 with the statistical packet SNPassoc to obtain three-dimensional graphics. Cross-validation analysis and estimation of Individual SNP informativeness (I_n) was performed using the “Snipper app suite” classifier <http://mathgene.usc.es/snipper/>.

Results

Characterization of PIMA-AIMs

We selected 27 AIM-SNPs complementing genomic positions to those occupied by the markers previously established in a 34-plex assay to detect differences among four major population groups (Sub-Saharan Africans, Europeans, East Asians and Native-Americans). Allele frequencies, chromosomal location, F_{st}/I_n values, rs numbers and primer sequences are shown in **Supplementary Table S1**. All markers were in HWE and pairwise linkage disequilibrium in each ancestral group. A comparison of principal component analysis (PCA) of the reference populations using both PIMA and 34-plex sets combined indicated that adding the PIMA SNPs to the 34-plex panel, provided a significant degree of extra resolution to give complete separation of the closely related American and East Asian groups. This is shown in the PCA plots on **figure 2**, where American and East Asian samples occupy almost identical positions using the 34-plex SNPs, the Americans are most distantly positioned using PIMA, but the most balanced separations of each of the four groups is achieved with the combined marker sets.

Ancestry inference accuracy of small AIM-panels (PIMA/34plex) in comparison with larger AIM-panel.

Ancestry inference accuracy of PIMA/34plex simultaneous analysis was compared to a larger panel recently published (LACE panel) comprising 308 AIM-SNPs specific for detecting Native-American component in three-hybrid populations (Native-

American, African and European ancestry) (Galanter, 2012). **Figure 3** represents the Structure cluster plots comparing both panels sets; at the top PIMA/34plex and at the bottom LACE. Assumed cluster $K=3$ and $K=4$ are depicted for both panels. Although estimation of East Asia ancestry was not a selection criteria in the LACE panel, the reference population from this geographic group was equally well differentiated from the other populations when analyzing global pattern of variability amongst the AIM-SNPs assuming four major clusters. The individual estimated ancestry across the study populations, assuming either three or four-population models, showed similar patterns of admixture in both panels. Generally, there was observed a similar relation between F_{st} in both panels, with lower values for East Asian population, followed by Europe, America and Africa. For informativeness values (I_n), a slight variation between the panels was observed, showing the highest value for America in Pima+34plex, while for LACE panel the major value of informativeness was represented by African population. Population assignment through cross validations was estimated with Snipper app suite. Both panels showed a complete differentiation (100% classification success in all population) for the optimum population assumed ($K:4$) from references population analyzed: Africa, Europa, East Asia and America. However assuming $K:3$ there is a slight underestimation of Asiatic population observed in the LACE panel in comparison with Pima/34plex set. In relation with the admixed population compared in both panels the correlation (r^2) in the ancestral component are represented in **Supplementary Figure 1**

Patterns of admixture in populations from North, Central, South America and the Caribbean

Admixed patterns of North, Central/ South American and the Caribbean study populations showed consistent results with both the geographic distribution and the demographic histories. **Figure 4** shows the Structure cluster plot for all populations analyzed, placed alongside a raster plot which summarizes the patterns of contrasting allele frequencies, notably amongst the fixed difference SNPs providing the most informative markers. The raster plot, where the informative homozygotes are red bands (opposite homozygotes green and heterozygotes grey) gives a complete perspective of the component SNPs allele distributions and where the greatest differentiation exists for each marker. With the exception of Colombian

Mulalo, Brazilian Riacho de Sacutiaba and Kaluga, and de ASW from north America, which presenting a predominant African ancestry as expected, the urban populations consistently showed majority European ancestry while the tribal or rural populations tended to show majority Native-American ancestry. The sample set from Guatemala City more closely matched the tribal sample from this region having predominantly Native-American ancestry. Notably, the Colombian Awa was the only study population to show a high patterns of Native American ancestry comparable to those of the American CEPH reference populations.

In view of such a wide range of admixture patterns observed in the study populations, the best approach to the data is to compare populations with similar admixture components and ratios. The populations were grouped into triangles plots shown in **Supplementary figure 2**, with identical African, European and American component population vertices and reference clusters in each case. These comprise by geographic areas. Briefly summarizing each plot in turn: Guatemalans show a smooth gradient between full American ancestry and equal AME-EUR components; with Colombian, Venezuelan and Chilean patterns largely matching this one but revealing a discernible African (AFR) component in several Colombians and most Venezuelans, in line with their proximity to the Caribbean. As expected, Brazilians from Sao Paulo show the greatest admixture heterogeneity with a full gradient of EUR-AFR-AME component ratios and Kaluga and Riacho de Sacutiaba with an african gradient, also Colombian Mulalo showing AFR ancestry rather than EUR to be the predominant component. The seven Native American populations show some stratification of component ratios ranging from the fully AME ancestry of the Awa to equal AME-EUR components in Colombian Pijao. Finally, it is interesting to observe that in the last plot (admixed), even more stratification of component ratios with each admixed study population showing largely non-overlapping positions in the triangle reflecting differing degrees of influence of African and European admixture on Native American SNP variability. The Afro American population (ASW) showed mainly African component followed by a European component, the Mexico and Puerto Ricans populations presented a high European ancestry, followed by a Native and African component respectively.

Comparison of autosomal AIM-SNPs *versus* lineages markers in admixed population from Colombia and Venezuela

Supplementary figure 3 shows graphic bars comparisons (either at population or individual level) for admixed Colombian and Venezuelan samples. PIMA/34-plex membership estimated ancestries are placed alongside to mtDNA and Y chromosome haplogroups. The overall picture confirms previous findings of biased mating between men of European ancestry with women from all ancestries mainly Natives and African. The autosomal markers were able to detect a much larger European component in front to mtDNA, placing male ancestors, as the mainly European source. On the other hand, the Native-American contribution detected by mtDNA was higher in front to autosomal markers. Mulalo was a characteristic population, where the African component was predominant over the others specially regarding mtDNA haplogroups. Data from Y chromosome haplogroups were also available from Awa, NW Colombia and Mulalo. Comparing these data to individual analysis in Venezuelan population confirm the scenario of biased mating and reaffirms the need to analyze autosomal markers in order to understand the contributing population as a whole and have a wider description from an individual ancestry.

Discussion

In the present work we have demonstrated that a small battery of SNPs owns the capability of accurately predict the genetic structure of admixed American population as were demonstrated after comparison with a wider panel set. The results in this SNP panel provide a direct reflection of the complex patterns of admixture in Native American Populations membership and standard error.

The results shows that despite centuries of inter-ethnic mating between people from different continents, the Native-American substrate persists in the present-day as were observed after markers comparison in Colombian and Venezuelan populations. But it is also remarkable that an European and African component are present. For example, Mulalo population group located in central-western Colombia, with moderate geographic isolation, has been the subject of many studies, ethnic, sociologic and genetics our results confirm the great African component (58%) kept since 1800 after initial era of slavery in South America, although it shows the mix of European (24%), Native American (17%) and Asian (1%). The Wayu group was one of the most extensive in the region now known as Zulia State in Venezuela for the arrival of the Spanish, was a highly eroded and is quickly assimilated into the process, which resulted in an important mix with the rest of the populations in the

Spanish but retain strong indigenous culture is seen strongly admixed process them over the years. One of our most relevant hypothesis was whether the Brazilian population had a large Native American component, but it is clear that at least the population of Sao Paulo is only mixed African (25.2%) and European (60.2%) and only (13.9%) Native American component and only (6%) Asian component, it would be necessary to examine other populations to the north of Brazil the Amazon jungle to define a component major. In addition the *Quilombos* from Brazil are also an example of the different pattern of admixture in the region which corresponds to their demographic history.

In relation to the comparative analysis between lineages and autosomic markers, the data from Carvajal et al. indicate that »94% of the Y chromosomes are European, 5% are African, and 1% are American. Y-chromosome data are consistent with an origin of founders predominantly in southern Spain but also suggest that a fraction came from northern Iberia and that some possibly had a Sephardic origin. In stark contrast with the Y-chromosome, »90% of the mtDNA gene pool of Antioquia is American, with the frequency of the four American founder lineages being closest to Native Americans currently living in the area. These results indicate a highly asymmetric pattern of mating in early Antioquia, involving mostly immigrant men and local native women. The discordance of our data with blood-group estimates of admixture suggests that the number of founder men was larger than that of women. Our results confirm this information because show that this population is 11% African, 54% European and 34% Native American. The Awa people of Colombia can be considered as almost not admixed in comparison with other American populations of the CEPH panel (Pima, Surui and karitian) which coincides with the information obtained with markers of the Y chromosome and mitochondrial DNA in the present work.

As a final remark, the PIMA together with our previously developed AIM panel 34-plex detect the difference between the American population and the other 3 main geographic groups involved in admixture process in modern population in the Americas, with a comparative classification power respect to a wider panel recently published. In addition, this panel show a stronger differentiation between American and Asiatic population. These advantages (reduced number of SNP and high resolution in differentiation) make this panel and valuable tool for the practical implementation in population genetics studies.

References

- [1] Alvarez-Iglesias V, A. Carracedo, A., Salas (2007). Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. *Forensic Sci Int Genet* 1, 44-45.
- [2] Alvarez-Iglesias V, M.-M. A., Cerezo M, Quintans B, Zarrabeitia MT, Cusco I, Lareu MV, García O, Perez-Jurado L, Carracedo A, Salas A (2009). New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. *PLoS One* 4, e5112.
- [3] Achilli, A., Perego, U. A., Bravi, C. M., Coble, M. D., Kong, Q. P., Woodward, S. R., Salas, A., Torroni, A., Bandelt, H. J. (2008). The Phylogeny of the four pan-American mtDNA haplogroups: implications for evolutionary and disease studies. *PLoS One* 3 (3), 1764.
- [4] Amigo, J., Salas, A., Phillips, C. (2011). ENGINES: exploring single nucleotide variation in entire human genomes. *BMC Bioinformatics* 12, 105.
- [5] Amigo, J., Salas, A., et al. (2008). SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics* 9.
- [6] Bedoya, G., Montoya, P., García, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., Hedrick, P., Ruiz-Linares, A. (2009). Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. *PNAS* 103(19), 7234-7239.
- [7] Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, et al (2002). A Human Genome Diversity Cell Line. *Science* 296, 261-262.
- [8] Carvajal-Carmona, L. G., et al (2000). Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Am J Hum Genet* 67(5), 1287-1295.
- [9] Earl, D. A. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*
- [10] Excoffier, L., Laval, G., Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online* 1, 47-50.
- [11] Frudakis, T. (2008). Molecular Photofitting. Predicting ancestry and phenotype using DNA. Inc., E., British Library Cataloguing-in-Publication Data.

- [12] Galanter, J. M. et al., (2012). Development of a Panel of Genome-wide Ancestry Informative Markers to Study Admixture Throughout the Americas. *PloS Genetics* 8, e1002554.
- [13] Gomez-Carballa, A., Ignacio-Veiga, A., Alvarez-Iglesias, V., Pastoriza-Mourelle, A., Ruiz, Y., Pineda, L., Carracedo, A., Salas, A. (2012). A melting pot of multicontinental mtDNA lineages in admixed Venezuelans. *Am J Phys Anthropol* 147 (1), 78-87.
- [14] Halder I, Shriver M, Thomas M, Fernandez J, Frudakis, T. (2008). A Panel of Ancestry Informative Markers for Estimating Individual Biogeographical Ancestry and Admixture From Four Continents: Utility and Applications. *Human Mutation* 29 (5), 648-658.
- [15] Jakobsson, M., Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23(14), 1801-1806.
- [16] Li, J. Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-1104.
- [17] Brion, M., Sobrino, B., Blanco-Verea, A., M. V. Lareu, Carracedo, A. (2004). Hierarchical analysis of 30 Y-chromosome SNPs in European populations. *Int J Legal Med* 119, 10-15.
- [18] Malamud, C. (1993). *Historia De América: (Temas Didácticos)* Editorial Universitas).
- [19] Mao, X., Bigham, A., Mei R., Gutierrez, G., Weiss, K., Brutsaert, T., Leon-Velarde, F., Moore, L., Vargas, E., McKeigue, P., Shriver, M., Parra, E. (2000). A Genomewide Admixture Mapping Panel for Hispanic/Latino Populations. *The American Journal of Human Genetics* 80, 1-7.
- [20] Pereira R, P. C., Pinto N, Santos C, Santos SEBd, et al. (2012). Straightforward Inference of Ancestry and Admixture Proportions through Ancestry- Informative Insertion Deletion Multiplexing. *PLoS ONE* 7(1), e29684.
- [21] Phillips, C., Salas, A., Sanchez, J. J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M. C., Ballard, D., Lareu, M. V., Carracedo, A. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet* 1(3-4), 273-280.

- [22] Phillips, C., Rodriguez, A., Mosquera-Miguel, A., Fondevila, M., Porras-Hurtado, L. R., F., Salas, A., Carracedo, A., Lareu, M. V. (2008). D9S1120, a *simple* STR with a common Native American-specific allele: forensic optimization, locus characterization and allele frequency studies. *Forensic Sci. Int. Genet* 7-13.
- [23] Pritchard, J., Donnelly, P. (2001). Case-Control Studies of Association in Structured or Admixed Populations. *Theoretical Population Biology* 60, 227-237.
- [24] Pritchard, J. K., Stephens, M., Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155(2), 945-959.
- [25] Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* 4(1), 137-138.
- [26] Rosenberg, N. A., L.M. Li, R. Ward, and J.K. Pritchard. (2001) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 73, 1402-1422.
- [27] Salas, A., Acosta, A., Alvarez-Iglesias, V., Cerezo, M., Phillips, C., Lareu, M. V., Carracedo, A. (2008). The mtDNA ancestry of admixed Colombian populations. *Am. J. Hum. Biol* 20, 584-591.
- [28] Salzano FM, MC, B. (2002). *The Evolution and Genetics of Latin American Populations*. Cambridge: Cambridge University Press
- [29] Shields, G. F., Schmiechen, A. M., Frazier, B. L., Redd, A., Voevoda, M. I., Reed, J. K., Ward, R. H. (1993). mt DNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. *Am. J. Hum. Genet* 53, 549-562.
- [30] Underhill, P. A., Jin, L., Zemans, R., Oefner, P. J., Cavalli-Sforza, L. L. (1996). A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc. Natl. Acad. Sci* 93, 196-200.
- [31] Volodko, N. V. (2008). Ome diversity in arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas. *Am J Hum Genet* 82, 1084-1100.



Figure 1: Geographic location from all study and reference population from America.

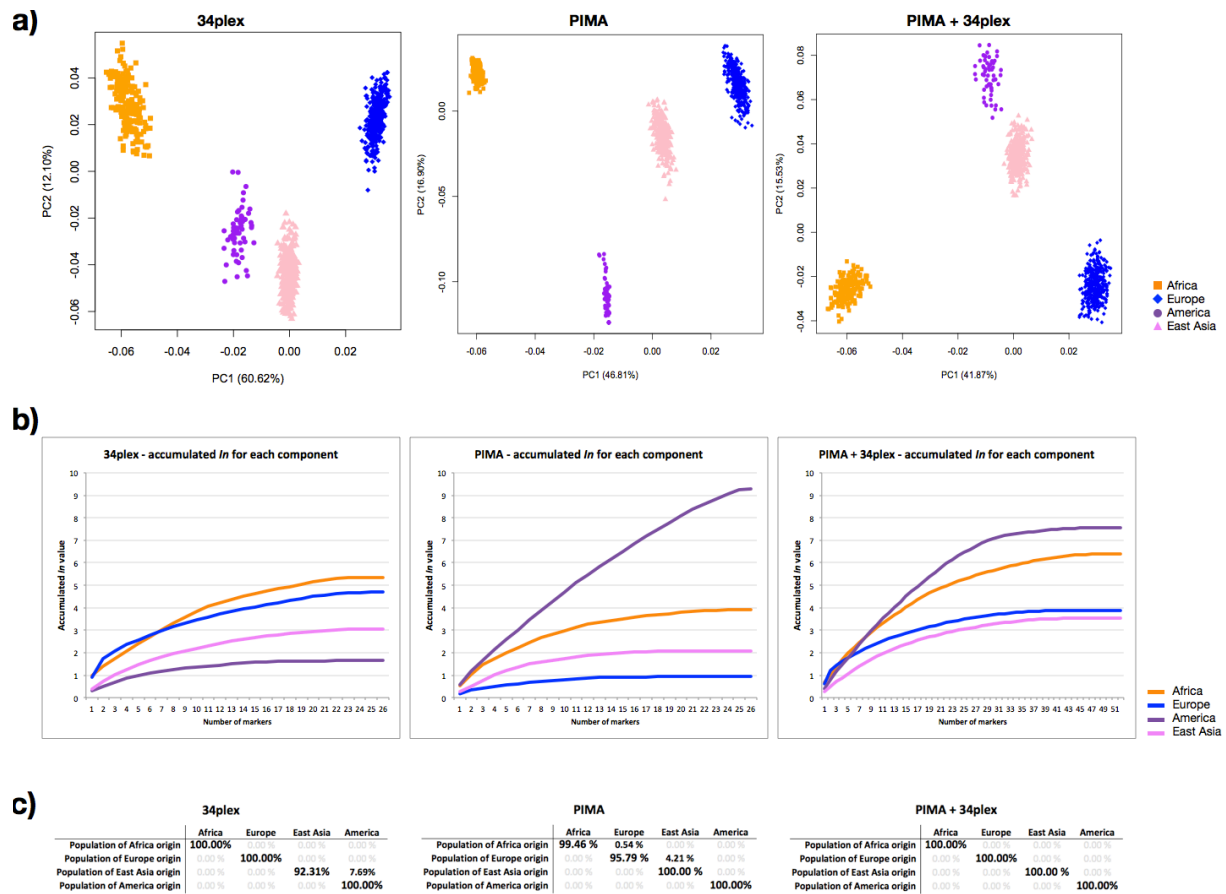


Figure 2: Comparison of principal component analysis (PCA) of the reference populations using both PIMA and 34/plex.

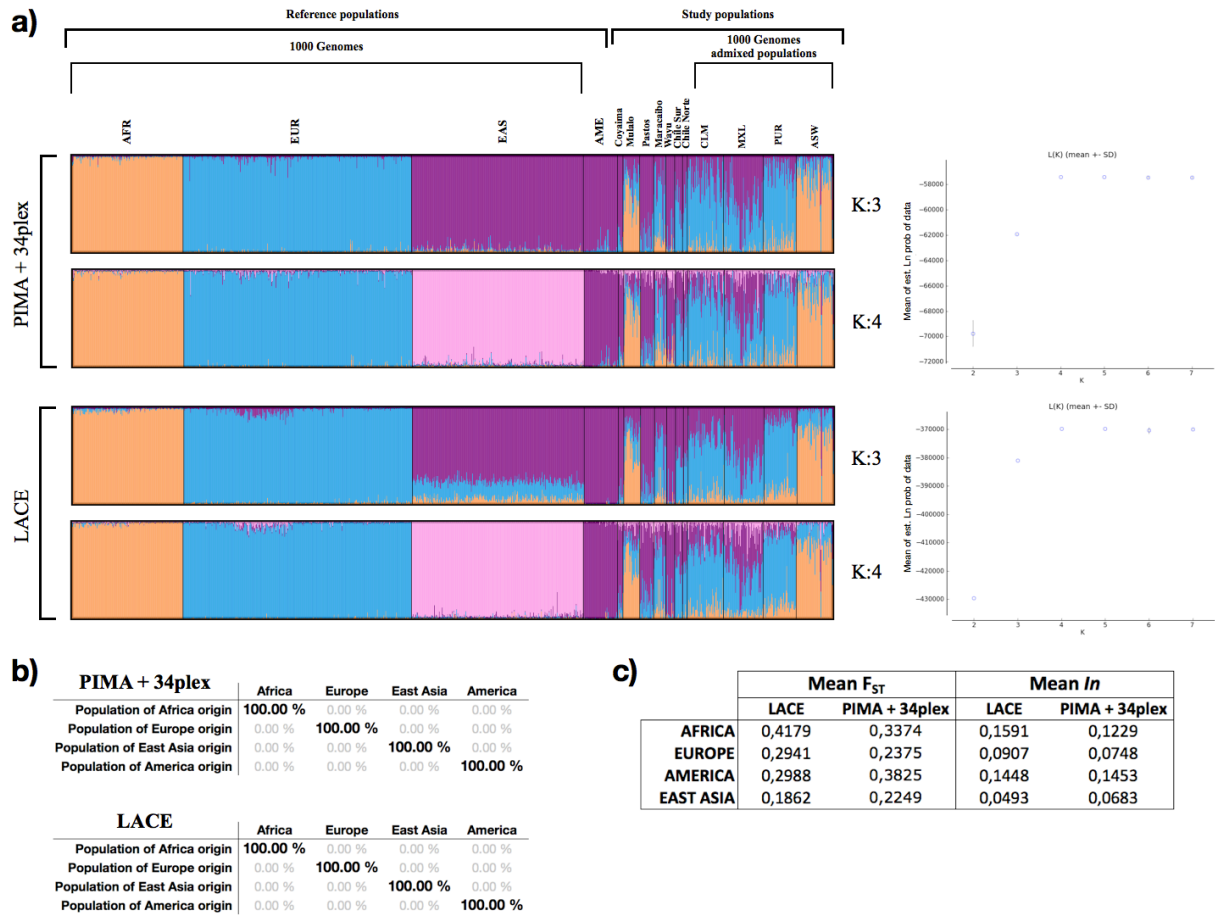


Figure 3: Structure cluster plots and classification success with Snipper, comparing LACE and PIMA/34Plex panels.

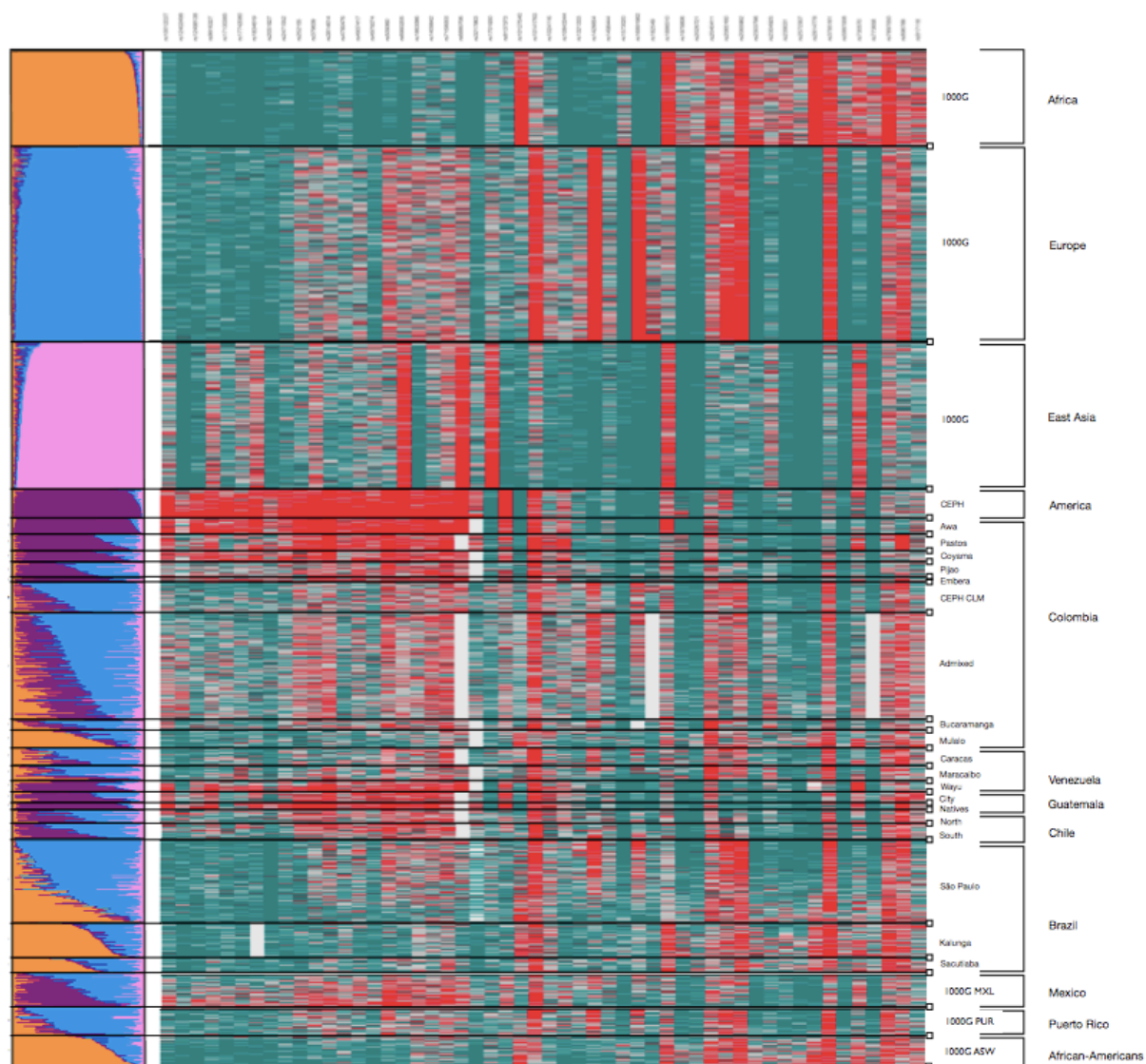
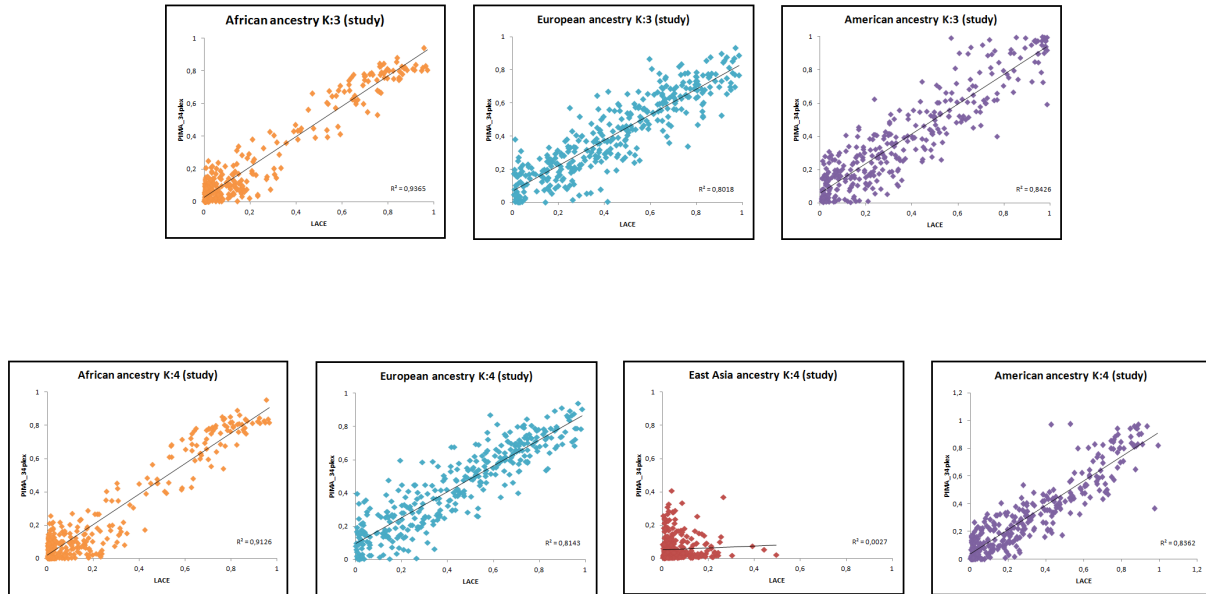
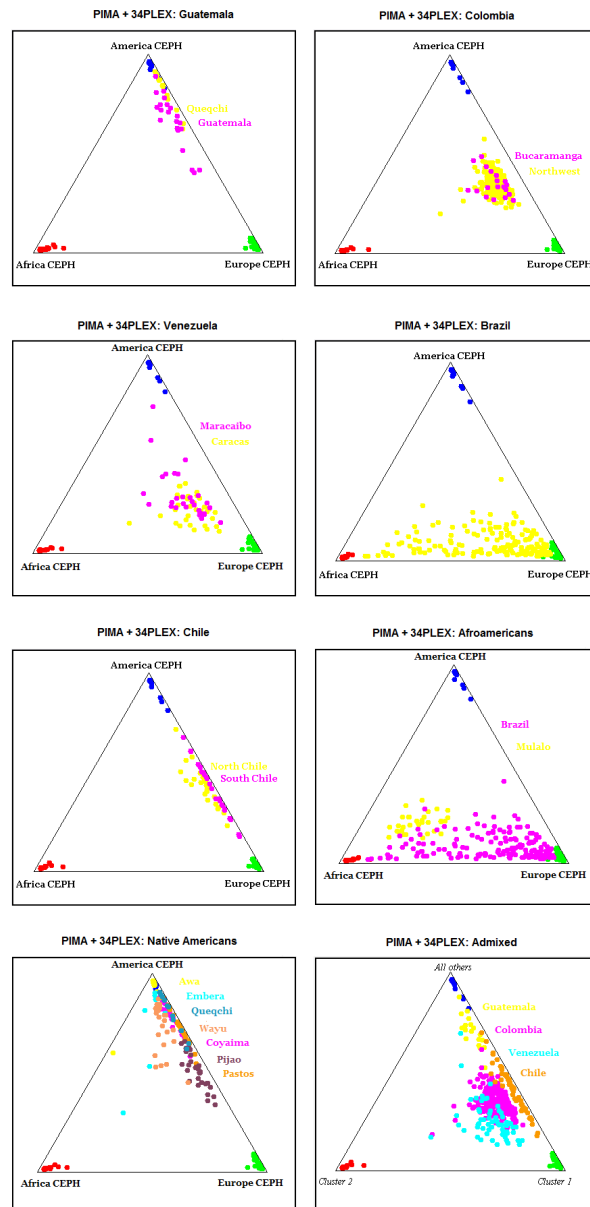


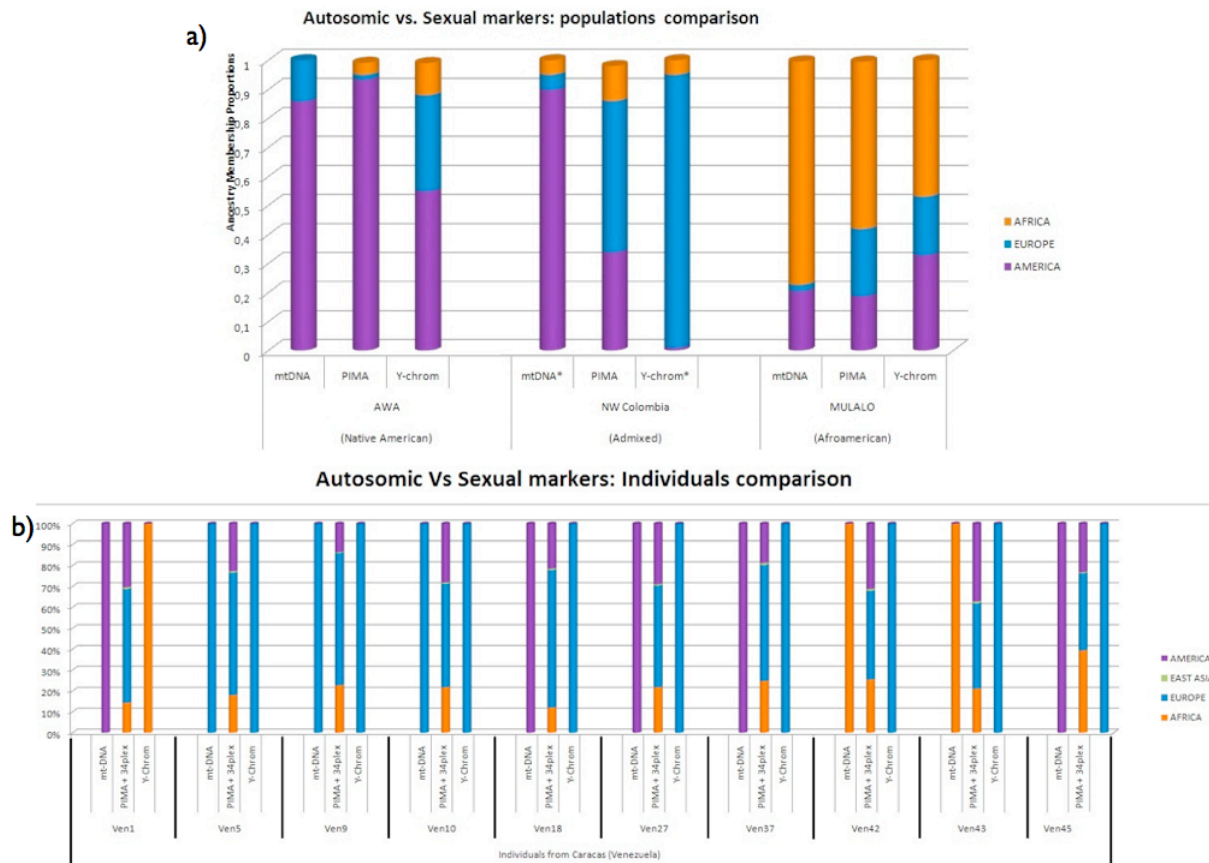
Figure 4: Structure cluster plot for populations study analyzed placed alongside a raster plot.



Supplementary Figure 1. Analysis of linear regression between ancestral components in study population compared in Lace and pima/34plex.



Supplementary Figure 2: Triangles plots from Structure representing the analysis of study populations grouped by geographic location or admixture characteristics compared to the major admixture components.



Supplementary Figure 3: Comparative analysis between lineages (mtDNA and Y chromosome) and autosomic markers (Pima+ 34plex) in population from Colombia (a) and individuals from Venezuela (b).

Supplementary Table S1: Allele frequencies, chromosomal location, Fst/In values, rs numbers and primer sequences.

Bloque 1- Discusión

La historia demográfica del continente americano resulta de gran interés, no sólo debido a la composición multi-étnica de sus poblaciones urbanas, sino también por la presencia de grupos nativos poco caracterizados, los cuales preservan un *pool* genético reflejado en los patrones de divergencia entre éstas y otras poblaciones, como lo indican los estudios realizados en el presente trabajo de investigación. Tal es el caso de las poblaciones nativas de Venezuela pertenecientes a la región de Guayana (Panare, Warao y Pemón), las cuales no habían sido caracterizadas genéticamente hasta la fecha con propósitos forenses y de estudios de grupos ancestrales. De igual forma, existen aproximadamente otras 25 etnias en esta región no caracterizadas genéticamente, por lo que resulta de interés dada la diversidad étnica existente, expandir los estudios que permitan la exploración de nuevos grupos filogenéticos, y que además contribuyan a la comprensión sobre la composición de las poblaciones actuales de la región. La presencia de una baja diversidad genética (haplotípica y nucleotídica) en el componente nativo, y altos valores de *Random Match Probability* en estas etnias en comparación con otros grupos, sugiere un patrón genético “atípico”. Esta diferenciación puede ser debida en parte, a un estado de aislamiento, posiblemente definido por barreras geográficas (como el río Orinoco y el Macizo Guayanés). Sin embargo, aunque existe una evidente diferenciación entre éstas y otras poblaciones, señalada por las frecuencias alélicas observadas en SNPs autosómicos (tanto en AIMs como de identificación individual), su composición ancestral no es únicamente Nativoamericana (~98,9%), lo que indica la existencia de una cierta mezcla. Los estudios de ADN mitocondrial y de cromosoma Y que se encuentran en curso, sugieren cierto “sesgo por el sexo” en esta mezcla, ya que para las mismas poblaciones nativas un 100% de sus linajes mitocondriales son amerindios, mientras que por estudios de cromosoma Y, un 95% de sus linajes son amerindios, un 4% europeos y un 1% africanos (Figura D1).

El comparar los análisis obtenidos de marcadores de linajes frente a AIM-autosómicos, aporta aún más validez los paneles de marcadores de grupos ancestrales presentado. Esto debido a que nos permiten tener una visión más completa sobre el aporte por parte de poblaciones “fundadoras” respecto a una mezcla más reciente en la historia de estas poblaciones.

En la presente investigación ha sido posible realizar nuevos aportes de interés sobre la diversidad de estas poblaciones multi étnicas. La variabilidad en la distribución genética de los tres principales grupos ancestrales (América, Europa y África) que conforman a estas poblaciones del continente, guardan una estrecha relación con los datos demográficos que han sido registrados históricamente. Poblaciones predominantemente afro-descendiente, como los Chocó, Mulaló y Yungas, corresponden a localidades donde se desarrollaron grandes focos de tráfico de esclavos africanos durante la conquista del continente. Otras poblaciones estudiadas en Argentina, Chile y algunos grupos urbanos de Colombia y Venezuela, presentan un componente predominantemente europeo, seguido por el nativoamericano y africano. Estos grupos urbanos presentan un grado de mezcla mayor y más diverso en comparación con otros, producto de las distintas migraciones que continúan ocurriendo dentro, y desde fuera del continente. Por lo tanto, estos grupos urbanos corresponden a poblaciones que no han estado sometidos a un tipo de aislamiento, a diferencia de aquellos con ascendencia predominantemente nativa, los cuales representan un recurso importante para el estudio de la variabilidad en el continente. Por lo general, las poblaciones de gran diversidad en la región de estudio suelen estar representadas por grandes ciudades o capitales. La ciudad de Caracas (Venezuela), es un ejemplo de un grupo de población multi-étnica, que presenta una diversidad mayor que otras localidades del país, como por ejemplo Pueblo Llano, ciudad al interior del país cercada por la cordillera de los Andes, con una composición nativa mayor que Caracas, mientras que con una influencia de grupos ancestrales europeos y africanos menor en comparación con la capital. Si comparamos además una población caraqueña (un sub-conjunto correspondiente a los datos genotipados en V.3) con la nativa de Venezuela (analizada en V.2), la diferencia en cuanto a su diversidad se hace aún más evidente, observando también un aporte nativo exclusivamente mitocondrial, lo que confirma el patrón de sesgo observado previamente en el grupo nativo (Figura D1), así como una mayor diversidad de haplogrupos del cromosoma Y incorporados al continente, en comparación con los grupos mitocondriales.

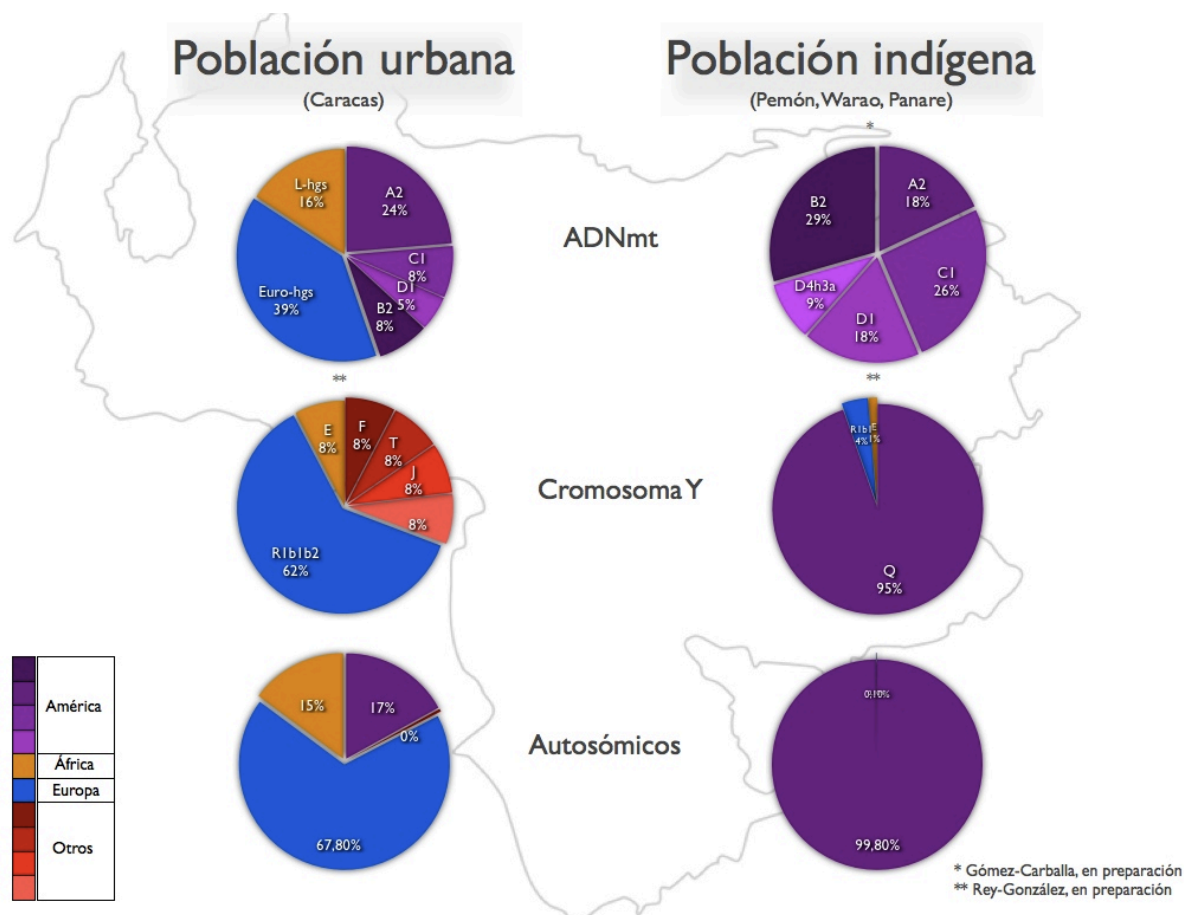


Figura D1. Comparación de marcadores autosómicos, ADNmt, y cromosoma Y en poblaciones nativas y urbanas de Venezuela.

Además de los tres principales componentes que han contribuido a la composición de la gran mayoría de las poblaciones actuales en el continente americano, existe también un aporte más reciente de otros grupos geográficos provenientes de Asia y Oriente Medio, los cuales aunque en menor medida, representan también poblaciones de interés en el estudio genético y demográfico del continente. Tal es el caso de los eventos de inmigración originarios de Japón, China y Taiwan que ocurrieron mayormente en Brasil, Perú y Chile a finales del siglo XIX y principios del siglo XX, producto de una crisis demográfica en estos países y de una creciente demanda de mano de obra en América. Por otra parte, algunos ejemplos de inmigración proveniente de Oriente Medio en América son los originarios de Palestina, Líbano y Siria, grupos dedicados principalmente al comercio en países como Argentina, Chile y Venezuela desde comienzos del siglo XX. En la figura D1 están representados haplotipos del cromosoma Y de algunos de estos grupos minoritarios en un muestreo realizado en la ciudad de Caracas (trabajo en curso).

Por lo tanto, el empleo de paneles de marcadores genéticos que incluyan la componente aportada por otros grupos continentales además de los nativos, europeos y africanos, representan herramientas de utilidad en el estudio genético de poblaciones multi-étnicas en el continente americano, y cuyo impacto en la detección de mezcla genética será cada vez mayor con el tiempo. Paneles como los empleados en los estudios V.4 y V.6, representan una alternativa y punto de partida de interés para diferenciar estos componentes, además de concentrar un bajo número de SNPs, pero de elevado poder de información y de utilidad en el análisis forense de muestras difíciles (*challenging DNA*).

Finalmente, es importante establecer en todo panel informativo de grupos ancestrales un equilibrio adecuado durante la selección de marcadores. Un desequilibrio podría originar sesgo por aquellos marcadores que definieran mejor un grupo étnico en particular. En general todos los paneles de AIMs, incluyendo los presentados en este trabajo, son vulnerables a cierto grado de sesgo. Sin embargo, resulta conveniente minimizar el impacto de este sesgo a través de herramientas como las que fueron presentadas en el presente estudio. Por otra parte, cabe destacar que los paneles presentados fueron diseñados específicamente para detectar patrones de estructura a nivel intercontinental, por lo que en el caso de que no se detectara estratificación alguna tras su empleo, sería conveniente confirmar la ausencia de subestructura con otros *sets* de marcadores de diferenciación específica a nivel intra continental.

Bloque 2.

V.5. Further development of forensic eye colour predictive tests

Y. Ruiz, C. Phillips , A. Gomez-Tato, J. Alvarez-Dios, M. Casares de Cal,
R. Cruz , O. Maroñas, J. Söchtig, M. Fondevila, M. J. Rodriguez-Cid,
Á. Carracedo, M.V. Lareu

(Forensic Science International:Genetics, 2012, doi:10.1016/j.fsigen.2012.05.009)



Contents lists available at [SciVerse ScienceDirect](#)

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig



Further development of forensic eye color predictive tests

Y. Ruiz^a, C. Phillips^{a,*}, A. Gomez-Tato^b, J. Alvarez-Dios^b, M. Casares de Cal^b, R. Cruz^c, O. Maroñas^a, J. Söchtig^a, M. Fondevila^a, M.J. Rodriguez-Cid^d, Á. Carracedo^{a,c}, M.V. Lareu^a

^a Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Spain

^b Faculty of Mathematics, University of Santiago de Compostela, Spain

^c CIBERER, Genomic Medicine Group, University of Santiago de Compostela, Spain

^d Ophthalmology Department, Clinical Hospital, Santiago de Compostela, Spain

ARTICLE INFO

Article history:

Received 12 October 2011

Received in revised form 20 April 2012

Accepted 17 May 2012

Keywords:

Eye color

Pigmentation

Forensic phenotyping

SNPs

Classification algorithms

Structure

Bayesian classification

ROC

AUC

ABSTRACT

In forensic analysis predictive tests for external visible characteristics (or EVCs), including inference of iris color, represent a potentially useful tool to guide criminal investigations. Two recent studies, both focused on forensic testing, have analyzed single nucleotide polymorphism (SNP) genotypes underlying common eye color variation (Mengel-From et al., *Forensic Sci. Int. Genet.* 4:323 and Walsh et al., *Forensic Sci. Int. Genet.* 5:170). Each study arrived at different recommendations for eye color predictive tests aiming to type the most closely associated SNPs, although both confirmed rs12913832 in *HERC2* as the key predictor, widely recognized as the most strongly associated marker with blue and brown iris colors. Differences between these two studies in identification of other eye color predictors may partly arise from varying approaches to assigning phenotypes, notably those not unequivocally blue or dark brown and therefore occupying an intermediate iris color continuum. We have developed two single base extension assays typing 37 SNPs in pigmentation-associated genes to study SNP-genotype based prediction of eye, skin, and hair color variation. These assays were used to test the performance of different sets of eye color predictors in 416 subjects from six populations of north and south Europe. The presence of a complex and continuous range of intermediate phenotypes distinct from blue and brown eye colors was confirmed by establishing eye color populations compared to genetic clusters defined using Structure software. Our study explored the effect of an expanded SNP combination beyond six markers has on the ability to predict eye color in a forensic test without extending the SNP assay excessively – thus maintaining a balance between the test's predictive value and an ability to reliably type challenging DNA with a multiplex of manageable size. Our evaluation used AUC analysis (area under the receiver operating characteristic curves) and naïve Bayesian likelihood-based classification approaches. To provide flexibility in SNP-based eye color predictive tests in forensic applications we modified an online Bayesian classifier, originally developed for genetic ancestry analysis, to provide a straightforward system to assign eye color likelihoods from a SNP profile combining additional informative markers from the predictors analyzed by our study plus those of Walsh and Mengel-From. Two advantages of the online classifier is the ability to submit incomplete SNP profiles, a common occurrence when typing challenging DNA, and the ability to handle physically linked SNPs showing independent effect, by allowing the user to input frequencies from SNP pairs or larger combinations. This system was used to include the submission of frequency data for the SNP pair rs12913832 and rs1129038; indicated by our study to be the two SNPs most closely associated to eye color.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The visible physical characteristics of an individual form the main description of an eyewitness's testimony [1]. Recently,

considerable interest has been expressed in the viability of using genetic predictors of physical characteristics when reliable eyewitness is not available to police investigations or no national DNA database entry exists. The initial goal has been the design of predictive tests for common pigmentation variation, the first of which have typed coding region single nucleotide polymorphisms (SNPs) providing inference of iris color [2].

Eye color variability is considered to arise from variation in the number and distribution of stromal melanocytes and the melanosomes they contain [3,4]. In dark eye colors higher levels of

* Corresponding author at: Institute of Forensic Sciences "Luis Concheiro", University of Santiago de Compostela, Rúa San Francisco, s/n, 15782 Santiago de Compostela, La Coruña, Spain. Tel.: +34 881812311; fax: +34 981580336.
E-mail address: c.phillips@mac.com (C. Phillips).

melanin absorb more light, while an absence of this pigment in the stroma tends to disperse light and absorbs the majority of the color except blue and blue-gray. Complex intermediate tones, including green and hazel result from diverse quantities of melanin that reflect light of mixed tonalities [5]. Furthermore, different shades of these intermediates tones are determined by the thickness and density of the iris itself [6]. Methods for iris phenotype classification based on quantitative analysis through digital photography have been recently proposed, estimating values of saturation and tonalities [7] or using color charts [8], but the reliable characterization and assignment of intermediate iris color phenotypes remains very difficult and somewhat subjective [9].

Early eye color association studies identified OCA2 as the gene sited in the most closely associated region [9–12], but these were followed by the simultaneous identification, from three independent studies in 2008 [13–15], of the neighboring HERC2 gene, specifically the SNP rs12913832 [13,14], as the key human eye color regulator. A model for brown and blue determination was proposed based on regulation of OCA2 expression by promotor SNPs, principally rs12913832, embedded in HERC2 [6,14,16]. The predictive models originally proposed for OCA2 SNPs treated as haplotypes [9,12] were supplemented with analyses including HERC2 SNP haplotypes [6,16–18]. However following identification of rs12913832, the most comprehensive study of eye color associations in the HERC2–OCA2 region and in other pigmentation related genes was the landmark study of Liu et al. [19]. The main advantage of Liu's study was the systematic analysis of associations of very closely sited HERC2 and OCA2 SNPs, and whether these were due to close linkage with rs12913832 or independent effects. Liu also proposed a viable eye color prediction system using probabilistic analysis based on multinomial logistic regression that enabled the identification and listing of the strongest SNP predictors in order of effect. Notably, a very close SNP to rs12913832:rs7183877 (115 nucleotides apart) emerged as the 10th strongest predictor while the more distant rs1129038 (8759 nucleotides distant) was not identified as a predictor after adjustment for the effect of close linkage with rs12913832. Liu also identified SNPs in genes beyond the HERC2–OCA2 complex with predictive effect in TYRP1, TYR, SLC24A4, SLC45A2, IRF4, and ASIP [19].

Two smaller-scale SNP sets published in 2010 by Walsh et al. [20] (alternatively referred to here as *Irisplex* SNPs) and Mengel-From et al. [18] have adapted the above findings of Liu to provide forensic eye color tests. Each study arrived at slightly different recommendations for the development of SNaPshot single base extension assays typing just the most closely associated SNPs. Walsh used the six strongest SNP predictors from Liu, providing the most reliable inference of blue and brown iris phenotypes, but these predictors have less success predicting intermediate eye colors. Recently the detection of epistatic effects has provided a slight increase in the prediction of intermediate phenotypes [21,22], however further studies are needed to fully explain the genetic basis of these complex 'non-blue, non-brown' eye colors. Mengel-From identified four SNPs, all clustered around the HERC2–OCA2 complex, that were detected to be most closely associated when using a simplified light/dark iris phenotyping regime. The difference between the two studies in identification of best predictors may partly be due to their varying approaches in the assignment of intermediate iris colors, occupying a continuum that merges into blue and brown. However, the most significant difference between the studies was the use by Mengel-From of multiple SNP loci in the HERC2–OCA2 complex plus SLC45A2, whereas Walsh chose to implement a single SNP from each of six genes (the above three plus SLC24A4, TYR, and IRF4).

In our study, we examined the effect of bringing together the additional HERC2–OCA2 SNPs identified by Mengel-From with the

six *Irisplex* SNPs of Walsh. If an expanded SNP set provides improved predictions, adjusted for linkage, the total number of SNPs typed in a forensic test can still remain small, preserving the sensitivity and robustness of a SNaPshot-based assay. Prior to this current study we had developed two SNaPshot assays typing 37 SNPs in pigmentation-associated genes to analyze associations in skin, hair, and eye pigmentation variation in Europeans. We have used these assays to test the performance of different forensic eye color predictor combinations in 416 subjects taken from six populations of north and south Europe. Intermediate phenotypes, we defined as 'not simple blue or simple dark brown' iris colors, were compared to the genetic clusters obtained from *Structure* software. We adapted an online Bayesian classifier termed *Snipper* (<http://mathgene.usc.es/snipper/>), originally developed for ancestry analysis of genotype-based data [23] – to handle allele frequencies. This enhancement allows SNP-pair frequencies to be included in the data input by simple counting of each combination, although phase must be assumed. As *Snipper* handles user-defined custom SNP data, both as genotypes or frequencies, further predictor combinations could be accommodated, if identified in future. However submission of multiple-SNP haplotypes becomes complex and requires large sample counts to be made, therefore this study concentrated principally on the assessment of predictive performance when adding rs12913832–rs1129038 genotype pair data to the other SNPs of *Irisplex*, while identifying additional HERC2 SNPs as potential contributors to intermediate eye color variation.

Once we had identified an additional five SNPs on top of the six of Walsh and two of Mengel-From we made the 13 SNP genotype data available for use with *Snipper*. Beyond the rs12913832–rs1129038 HERC2 SNP pair, rs1667394 of HERC2 emerged as a potentially important interacting marker while rs7183877 appears to have a marked effect on prediction of green-hazel eye colors. Both these additional HERC2 SNPs were independently identified as significant eye color predictors by Eriksson et al. in 2010 [24].

2. Materials and methods

2.1. Population samples and phenotyping regimes

Study samples were collected from 416 volunteers from the following European populations: NW Spain (Galicia) 215; NW Germany (Lower Saxony) 91; Sweden (Dalarna) 44; Austria (Innsbruck) 31; Denmark (Copenhagen) 18 and Switzerland (Zurich) 17. Informed consent was obtained in all cases, as well as information about donor's immediate ancestry. Ethical approval was granted from the ethics committee of clinical investigation in Galicia, Spain (CEIC: 2009/246). DNA was obtained from buccal swabs that were stored at room temperature. DNA extraction was made using standard phenol–chloroform isoamyl alcohol protocols. Eye colors were recorded using a 12 megapixel reflex digital camera with uniform lighting conditions (consisting of a lens-attached ring flash), photographic settings and a macro lens that allowed the pupil, iris, and sclera to fill the majority of the image field. Phenotype assignment of eye color was defined following criteria described by previous studies [14,19] where "intermediates", i.e., non blue, non brown types were classified into green-hazel, intermediate-dark, and intermediate-light with the first category tending to have most examples of a brown peri-pupillary ring within a blue outer iris. However patterns for classification of intermediate phenotypes were also described according to the presence of melanin spots, accumulation of collagen, contract furrows and Fuch's crypts as detailed by Sturm and Larsson [6]. In order to compare our studies with those of Mengel-From, we added supplementary labeling of our study individuals into "light"

Table 1
SHEP 1 component SNP details.

ID	Gene	Position (Ref)	Chr	SNP	Primer forward	Primer reverse	Primer [μM]	Extension primer	Size (bp)	Probe [μM]	Sense	Original reporting publication(s)	Liu rank (top 15)
rs1042602	TYR	88551344	11	A/C	GGTGCTTCAATGGCAAAATC	TGACCTCTTTGTCTGGATGC	0.777	ttctctctctctcCAATGTCTCTCCAGATTCA	35	0.26	R	[31]	
rs26722	SLC45A2	33999627	5	C/T	TTTTTGCTCCGTGCAATGCC	GATCGAATGTACACGATATGG	0.518	ctctctctctctctTACGTAAACATTTTAACTTTCT	40	0.26	F	[10,32]	
rs12896399	SLC24A4	91843416	14	G/T	TCTGGCGATCCAATCTTTTG	GATCAGGAAGCTTAATCTGC	1.295	ttctctctctctctctctctcGGTCAGTATATTTGGG	43	0.47	R	[29,32]	3
rs11636232	HERC2	26060221	15	C/T	ACAGCAAAAGGGTCTGTTC	GCATTGAAGCGCCAAAAGTC	0.777	ctctctctctctctctctctctcagTGTCCCTCC-	47	0.38	F	[14]	
rs16891982	SLC45A2	33987450	5	C/G	TCTACGAAAGAGGAGTCGAG	AAAGTGAAGAAAAACACGGAG	0.777	ctctctctctctctctctctctctctcttgaGTTGGATGTT- GGGGCTT	49	0.28	F	[25-28]	4
rs13289	SLC45A2	33959166	5	C/G	GTGTTAAGTACCACGAGGAG	GTCACACCTTCTTCAAATC	0.907	ttctctctctctctctctctctctctctctctcGAGGAGAA- ATATCAGGGC	54	0.26	F	[26,28]	
rs7495174	OCA2	26017833	15	A/G	TAGGTCTGCGCTCCGTCCAC	GGCTTAGCAAGCAAGCGAAG	0.130	ttctctctctctctctctctctctctctctctcCAGGCCAAG- TTCCCTTAAGGT	56	0.09	R	[11,13,29]	8
rs1805007	MC1R	88513618	16	C/T	CTACATCTCCATCTTCTAC	ATGAAGAGCGTGTCTGAAGAC	0.907	ccccccctaaactagtgccacgcgtggaagctgacaa- CAGCATCTGCTGTAGC	60	0.26	R	[28,30]	
rs1667394	HERC2	26203777	15	A/G	ACAGCCAGCAATTCAAAACG	GAGACTTTGAGGTCTCCAAC	0.648	ttcTAG- CAATTCAAAACCTGCATA	64	0.28	R	[13-15]	9
rs1805008	MC1R	88513645	16	C/T	CTACATCTCATCTTCTAC	ATGAAGAGCGTGTCTGAAGAC	0.907	ccccctaaactagtgccacgcgtggaagctgacaa- GCCGCAACGGCTCGCCGCCCC	64	0.26	R	[10]	
rs916977	HERC2	26186959	15	A/G	TTCTGTCTCTCTTGACCCCG	GGTGTGGGATTGTGTTTGCC	0.130	ttc- ctcTAGCTTGGCAGCCTCT	71	0.14	F	[13-15]	
rs4778138	OCA2	26203777	15	A/G	CCTCCCATCACTGATTAGC	CAAACTCTCAAGGCAANTCAG	0.648	ccccccccccccccccctaaactagtgccacgcgtggaaa- gtctgacaaCTGATTAGCTGTCTCTG	72	0.26	R	[12,13]	
rs12203592	IRF4	341321	6	C/T	TTCAITCACTTTTGTTGGG	CATATGCTAAACCTGGC	0.907	ccccccccccccccccctacgaactaaactagtgccacgtc- gtgaaagtctgaCTGCTGCTGTAAAGAAAGG	76	0.26	F	[32]	6
rs12913832	HERC2	26009415	15	A/G	CGAGGCCAGTTTCATTGAG	AAAACAAAGACAGACCTCGG	0.130	ctc- ttAGCCAGTTTCATTTCAGCAITTA	76	0.14	F	[13,14]	1
rs3782974	DCT	93890897	13	A/T	ACCAAAAAATCAAAATCCAC	CCCATGATGATAAAATCCAATC	1.295	ttc- ctctctctATCCACTAATTTTGTGGAAGAG	80	0.47	F	[25,26]	
rs12592730	HERC2	26203954	15	A/G	AAGACAGAAAAGCTGCCAAG	GGATGCTTGAACAGATTATG	0.777	ttc- ctctctctcttGGATCCAATCAAAATTTACA	84	0.26	F	[14]	7
rs4778241	OCA2	26012308	15	A/C	AGGAGTGCAAATGTGTGGCTG	TGTACAGCCACTCTCGAAG	0.065	ttc- ttctctctctctctAAATTGTGGCTGTAGTCAATT	88	0.05	R	[12,15]	

Bold rs-numbers indicate the SNPs previously identified as most closely associated to eye color.

by grouping blue and intermediate light together plus “dark” by grouping brown with intermediate dark.

2.2. SNP selection, multiplex design and genotyping methods

A set of 37 SNPs considered to be associated with human pigmentation variation according to previously published studies was selected for genotyping [9–16,25–32] in two SNaPshot single base extension assays detailed below. Twenty-three of these SNPs found to be most associated with eye pigmentation in a series of published studies (literature references for all 37 SNPs are listed in Tables 1 and 2) were used to test the performance of forensic iris color prediction in the six European study populations. Tables 1 and 2 also list the ranking given by Liu for the eye color predictive power of the 15 most associated SNPs. The other 8 SNPs were additional HERC2 markers analyzed by the two studies that first identified rs12913832, plus components of a three SNP haplotype in OCA2 – which had been thoroughly analyzed by Kayser et al. [15]. Finally we included additional SNPs identified by three large-scale association studies of European pigmentation variation made by deCode [29,30,32].

Two novel single base extension multiplex assays were developed: SHEP 1 and SHEP 2 (i.e., skin, hair, and eye pigmentation) using Primer3 and AutoDimer [33,34] to design, check and optimize amplification primers creating amplicon lengths ranging from 87 to 135 base pairs (bp). Parallel extension primer sets were built in the same way and all primer sequences developed are given in Tables 1 and 2. Typical SHEP 1 and SHEP 2 primer extension profiles are shown in supplementary Fig. S1. Our aim when developing SHEP 1–2 was to examine all previously associated pigmentation SNPs and adapt these SNaPshot-based assays to smaller multiplexes later, when the most closely associated SNPs were more clearly identified. The alternative approach of designing Sequenom iPLEX assays lacked the necessary flexibility to modify SNP combinations with changing knowledge of the strongest associations or interactions.

The PCR reaction was optimized using 1–10 ng of DNA in 10 μ l final reaction volume, containing 1 \times Buffer, 1 \times BSA, 8 mM MgCl₂, 700 μ M dNTPs, 0.1–0.01 μ M of each primer and 0.1 U AmpliTaq Gold polymerase (Applied Biosystems, Foster City, US: AB). Amplification conditions comprised: denaturing at 95 °C for 10 min, then 35 cycles using 95 °C for 30 s, 60 °C for 50 s, 65 °C for 40 s, then a final extension of 65 °C for 6 min. Multiplexed SNaPshot (AB) single base extension chemistry was used to type the amplified SNP combinations in two parallel extension reactions. Prior to extension with SNaPshot 2.5 μ l of PCR product was treated with 1 μ l of ExoSAP-IT (USB[®] Corporation) to remove unused dNTPs or PCR primers, run at 37 °C for 15 min followed by 85 °C for 15 min to inactivate the enzyme. Then 1.5 μ l of purified PCR product was added to 2.5 μ l of SNaPshot ready reaction mix (AB) plus 1.5 μ l of extension primer mix (final concentration 0.2 μ M). Extension conditions comprised: 30 cycles of 96 °C for 10 s, 50 °C for 5 s, 60 °C for 30 s. The extension reaction products were cleaned up with 1 μ l of SAP (USB) at 37 °C for 80 min and 85 °C for 15 min. Capillary electrophoresis was performed on a Prism 3130xl Genetic Analyzer (AB) using Genemapper[®] Analysis Software v. 3.7 (AB).

2.3. Statistical analyses and classification models

One problem that occurs when comparing the association findings of Mengel-From with those of Walsh relates to use of light-dark and blue-intermediate-brown phenotyping regimes respectively. Because of this we decided to compare patterns of genetic clustering obtained with *Structure* [35] with our own assignment of five eye color phenotypes based on photography

(as described in Section 2.1), to assess how our subjects grouped, i.e., whether two, three or five groups are discernible from the genetic data of 23 SNPs. Analysis using *Structure* v. 2.3.3 comprised: 100,000 Markov Chain steps after a burn-in of length 100,000 with four replicates for each value of *K* (assumed populations) from 2 to 5. The admixture and linkage model was applied using LOCPRIOR information and frequencies correlated among populations. *CLUMPP* and *Distrupt* software were used to visualize and plot results [36]. We used standard approaches to calculate the optimum *K* value from the data based on the mean estimated probability of data stabilizing around a maximum value.

Each SNP was analyzed for Hardy Weinberg equilibrium (HWE) using 1,000,000 Markov Chain steps and assessed for pairwise linkage disequilibrium using “1000-permutation Chi-square tests” performed with Arlequin (v. 3.5). Significance analysis levels were adjusted for multiple tests following standard Bonferroni corrections [37].

Individual SNP informativeness as part of a predictive system was estimated using the *Snipper* classifier (<http://mathgene.usc.es/snipper/>), originally developed for genetic ancestry inference [20]. Informativeness was measured using Jensen and Shannon’s divergence for each marker [39] which is near-identical to Rosenberg’s informativeness for assignment (*I_n*) metric [40] that ranges from 0 = no divergence to 1 = maximum divergence. To evaluate the robustness of the eye color reference training sets for use with *Snipper* we performed two kinds of cross-validation: the classical one-out reclassification and a variant of the bootstrap analysis by randomly choosing (with replacement) a training set of 200 individuals from the reference set and classifying the remaining samples with this training set, repeating the procedure 100 times [38]. For this part of the assessment of training sets, a likelihood ratio threshold of 0.5 was used to denote a successful assignment. All calculations were made with custom programs written in R (v. 2.13.1).

As a Bayesian classifier based on likelihood ratios, *Snipper* sorts individual likelihoods in descending order then provides a prediction based on the ratio of the two largest likelihoods. We adapted the generation of likelihoods in *Snipper* to work with frequency-based training sets rather than genotypes (<http://mathgene.usc.es/snippet/frequencies.html>), allowing the frequencies of SNP combinations consisting of closely sited loci to be included in the training sets. The rs12913832–rs1129038 SNP pairs were counted without consideration of phase, e.g., AG (rs12913832) and AG (rs1129038) was treated as the profile component AA,GG, though strand1,strand2 can comprise AG,GA. To help readers assess this system we include three frequency-based training sets as supplementary Files S1–S3 for blue:green-hazel:brown eye colors: S1 that lists 23 individual SNP frequencies; S2 combining rs12913832–rs1129038 pairs with 22; and S3 combining rs12913832–rs1129038 plus the five other *Irisplex* SNPs. Smaller subsets of the 23-SNP data can be made by removing SNP worksheets accordingly, but SNP combinations beyond the rs12913832–rs1129038 have not been included as these are complex and difficult to count in sufficient numbers.

In order to allow a comparison with the *Irisplex* classification system we also calculated positive predictive values (PPV), negative predictive value (NPV), sensitivity and specificity values for the *Irisplex* 6, *Irisplex* 5 plus the rs12913832–rs1129038 pair; and *Irisplex* 5 + 2 plus HERC2 SNPs rs7183877 and rs1667394. Liu provides a systematic definition of the above values and the predictive model evaluation system used in the original study and applied to *Irisplex* in the supplementary data of [19]. Liu defines PPV as the percentage of correctly predicted color type among the predicted positives and NPV as the percentage of correctly predicted non-color type among the predicted negatives. Sensitivity is a measure of classification success, defined as: the percentage

ID	Gene	Position (Ref)	Chr	SNP	Primer forward	Primer reverse	Primer [μM]	Extension primer	Size	Probe [μM]	Sense	Original reporting publication(s)	Liu rank (top 15)
rs1015362	ASIP	32202273	20	AG	CCTTAAGTGTGTACTGTCTG	CTGAACAATAGTCCGACC	0.368	tctctctctcaTGTGTCTGAAACAGT	31	0.190	F	[16,29,30]	
rs1805005	MC1R	88513345	16	GT	AGGTGTCATCTCTGACGG	ACATGGGTGAGTCGACGTTC	0.674	ctctctctctctctcGTTGGAGAACGC- GCTGGTG	40	0.276	F	[10]	
rs7183877	HERC2	26039328	15	AC	GCCGAGGCTTCTTTTGTTT	CTGTCTCATGGCTACTAATC	0.490	tctctctctctctcAAGCAGTATACATTTA- GAATCGT	41	0.190	F	[15]	10
rs1408799	TYRP1	12662097	9	CT	TAGCACATGTCTGCTCGG	ATCAAACTGGTTCATCCAC	0.674	tctctctctctctctctctctctcCTCGGAGC- ACATGGTCA	46	0.207	R	[30]	12
rs1540771	IRF4-EXOC2	411033	6	AG	ATGGTAGAAGAGAGAGAGG	ACCACACGCTAGACATG	0.797	tctctctctctctctctctctTGAACTGC- ACGAGTTGG	46	0.241	R	[29]	
rs4911414	ASIP	32193105	20	GT	CCCCAGTCTCTTTTGTGTTG	GCGAACTAGAGAAAAACATC	0.245	ctctctctctctctctctctcGTCCTTGCT- GAGAAATTCATT	49	0.172	F	[30]	
rs1126809	TYR	88657609	11	AG	AATGGGTGCATTGGCTTCTG	CTCTGCAGTATTTTTCAGC	0.306	ctctctctctctctctctctctctctctcGAA- GAGCAGCGTGCCTT	53	0.190	R	[30]	
rs1393350	TYR	88650694	11	AG	GGAAGGTGAATGATAACAG	TACTCTTCTCAGTCCCTTC	0.306	ctctctctctctctctctctctctctctcAGT- CCCTTCTCGCAAC	53	0.069	F	[29]	5
rs4778232	OCA2	25955360	15	CT	AAGAACCAGGGATCTAGGG	CATGTCAGACTGTGAGATGG	0.429	ctctctctctctctctctctctctctctcG- GATCTAGGATGAGGAA	57	0.207	R	[15]	11
rs35264875	TPCN2	68602975	11	AT	CGTCTTCATTGTGTACTACC	CGTCAAAACAGTTGCTGGG	0.797	cccaactgactaaactagggtccagctgtggaag- taaaggGTGTACTCTCTTGGAG	60	0.310	F	[13,30]	
rs8024968	OCA2	25957284	15	AG	ACTTCACTTGTGTCCTTAG	TAGAGTCACAGAACAGGGAG	0.490	ctctctctctctctctctctctctctctctct- ctctaaTCCATAATCTCTTTCTCTGA	67	0.190	F	[15]	13
rs1800407	OCA2	25903913	15	AG	ATGATGATCATGCCCCACAC	ACTCTGCTTGTACTCTCTC	0.735	tctctctctctctctctctctctctctctct- tctctctctctctctctctctctctctctct-	71	0.310	R	[13]	2
rs1129038	HERC2	26030454	15	AG	CTTCTCATCAGACACACAG	TCGTGAGATCAGACGCTGAG	0.429	ctctctctctCATGGCCACACCCGTTCC	75	0.190	F	[11,13,14]	
rs1805009	MC1R	88514047	16	CG	TTTCTCGCCCTCATCATCTG	TCAGCACCTCTTTCAGCGTC	0.490	cccaactgactaaactagggtccagctgtggaag- aaactCTCTGCAATGCCATCATC	60	0.241	F	[29]	
rs6058017	ASIP	32320659	20	AG	AGCCGCCCTGTAGGGATCA	TCAGCCTCAACTGCTGAGCG	0.674	ctctctctctctctctctctctctctctctct- ctctctctctctctctctctctctctctctct-	79	0.241	F	[10]	14
rs6867641	SILC45A2	34021614	5	CT	AACGATCACACGCGTCTCT	GTAATAACGAGAAAAGCCCC	0.490	tctctctctctctctctctctctctctctctct- tctctctctctctctctctctctctctctct-	79	0.207	F	[27]	
rs1375164	OCA2	25965407	15	CT	ATAGGTACCTGTCTCTGTG	TAGAGGTCAATATCCCAGGGC	0.613	tctctctctctctctctctctctctctctctct- ctctctctctctctCTGCTGTGTGTGTA	83	0.241	R	[9,10,12]	
rs3829241	TPCN2	68611939	11	AG	TCCACAGGGATATCTGGAG	TGCTGGCTTCAGCGCTCTCTGT	0.368	tctctctctctctctctctctctctctctct- ctctctctctctctctCTGAGCTCATCTCC	83	0.190	R	[29]	
rs683	TYRP1	12699305	9	AC	CCACCTGTTGAATATAATAG	CCAGCTTTGAAAAGTATGCC	0.674	ctctctctctctctctctctctctctctctct- ctctctctctctctctctctctctctctct-	86	0.414	F	[10]	15
rs12821256	KITLG	87852466	12	CT	GTGAAGTTGTGTGGCAAG	TAAAGTTCCTCTGAGCCAAG	0.551	ctctctctctctctctctctctctctctctct- ctctctctctctctctctctctctctctct-	90	0.241	R	[29]	

Please cite this article in press as: Y. Ruiz, et al., Further development of forensic eye color predictive tests, *Forensic Sci. Int. Genet.* (2012), <http://dx.doi.org/10.1016/j.fsigen.2012.05.009>

of correctly predicted color type among the observed color type. Specificity is the percentage of correctly predicted non-color type among the observed non-color type.

IBM PASW SPSS Statistical-18 tests were used to analyze associations to phenotypes. Individual SNP associations were analyzed by logistic regression under an additive model. Adjustment for the most associated marker rs12913832 was made to detect the additional effect of other loci physically linked to this most strongly associated HERC2 SNP. We also made a SNP pair-adjusted analysis of the effect of other HERC2 SNPs adjusting for rs12913832–rs1129038 as a single variant and as two separate variants.

Following the classification approach used by Walsh et al. [20], we performed an analysis of AUC for ROC curves (area under the receiver operating characteristic curve) using the ROCR package [41], as a complementary method to assess the informativeness of three SNP sets: the six of *Irisplex*; these six plus two additional SNPs of Mengel-From, and; these 8 plus an additional five SNPs we identified as contributing detectable extra eye color predictability. All AUC analyses were made on the training set (256 samples) comprising: blue, brown, and green-hazel phenotypes.

Combined-effect and interactions between SNPs were analyzed using the multifactorial dimensionality reduction system (MDR) for pairwise eye color phenotype comparisons [42]. MDR is a non-parametric data mining approach that evaluates different

combinations of genetic or environmental factors (SNPs in our case). A set of n SNPs is selected (where n usually comprises 1, 2, 3, and 4 SNPs) and for each n SNPs and their possible multi-factorial classes (e.g., nine genotype combinations for 2 binary loci) the ratio of cases to controls is calculated. Each cell of a SNP combination is assigned to either a low- or high-risk group depending on the ratio of cases and controls. If this ratio meets or exceeds a threshold (usually 1.0) that genotype combination is determined as high risk, otherwise it is assigned as low risk. All potential combinations of n factors are evaluated sequentially and the model that gives the lowest error in classifying cases and controls based on low- or high-risk is selected for each set of n SNPs in a training set from 9/10 of the data, evaluating a test set from the remaining 1/10 of the data, to obtain the prediction accuracy. This cross-validation procedure, consisting of a random split of data into 9/10 and 1/10 proportions is repeated 10 times using a random seed number to protect against chance divisions of the dataset.

We applied the MDR software (www.epistasis.org) to determine the best 1, 2, 3, and 4 SNPs model for our dataset, producing four different models. Of the four best models, that with the highest average testing accuracy and cross-validation consistency (number of times a model is selected as best model among the validation sets) was selected as the final, best model. We used the MDR permutation module to test the significance of the association of this final model with case status. Finally, to

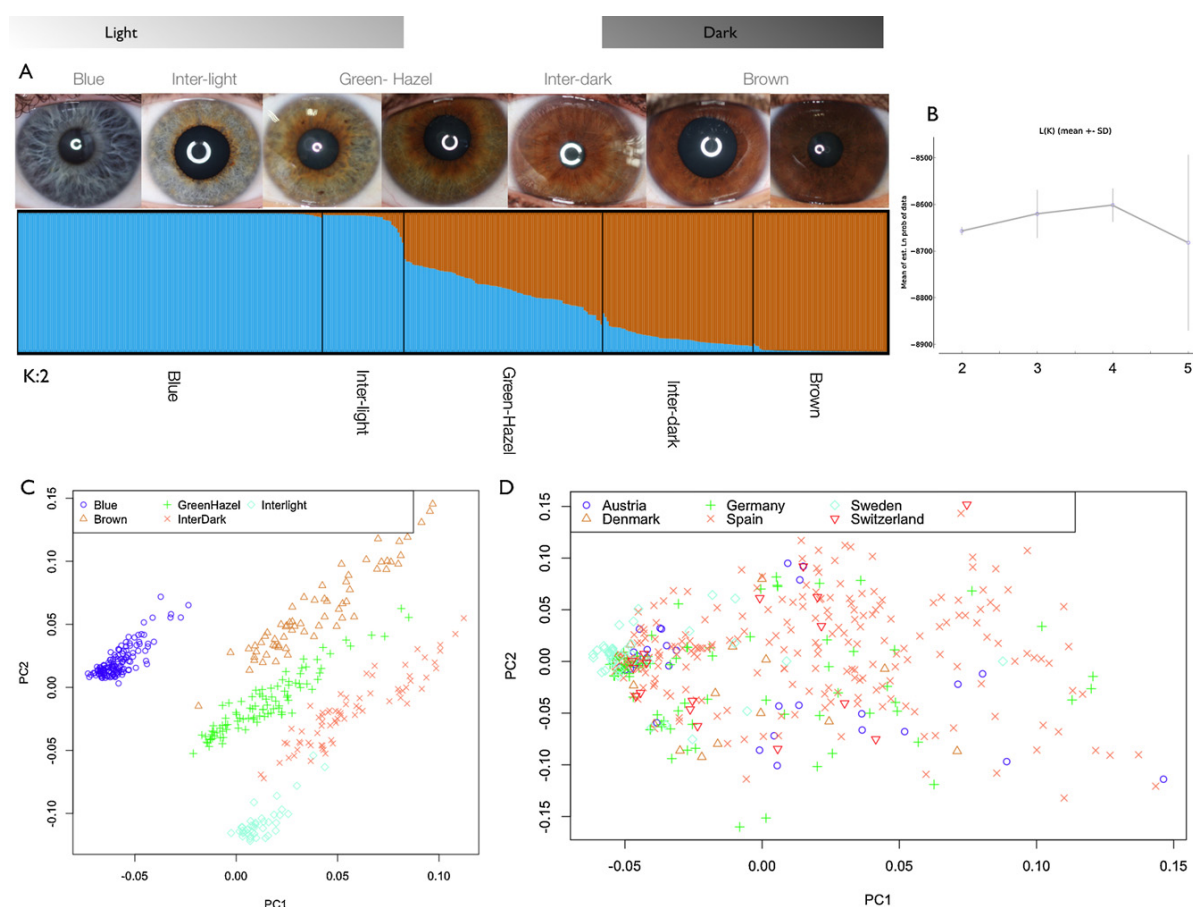


Fig. 1. (A) Analysis of eye color phenotype population structure in European samples with $K = 2$ clusters using *Structure* software. (B) Estimation of optimum K value indicates values greater than 2 can be inferred as the optimum cluster number. (C and D) PCA comparison of phenotype assignments made across population samples (C) and the geographical distribution of the populations studied (D).

visualize the relationships between the SNPs from the selected model we used the interaction dendrogram. This is constructed using a hierarchical cluster analysis and implemented in the MDR software. SNPs showing a strong interaction are located together and closely in the branches and the nature of the interaction is denoted by a color scale with blue indicating redundancy, to red, when SNPs show a high degree of synergy. MDR analyses were made separately for each color comparison (blue, inter-light, GH, inter-dark and brown vs. rest).

Lastly, principal component analysis (PCA) was made to compare the geographical distribution of phenotypes among our six sampled European regions, as populations in Europe tend to show a gradient of light to dark eye color running NW to SE. We used custom programs to execute PCA written in R (v. 2.13.1).

3. Results

3.1. Eye color phenotypes

Observed phenotypes across the European population samples consisted of 145 blue and 64 brown eye color plus 207 non-blue, non-brown phenotypes. The latter were not placed into a single group because of the full range of tonalities observed, so this group was classified as follows: green-hazel 95; intermediate-light 40; and intermediate-dark 72. We observed a high frequency of light eye colors in northern European subjects and this decreased moving south as shown in [supplementary Fig. S2](#), in agreement with previous observations [43,44].

[Supplementary Table S1](#) outlines the results of formal tests for Hardy Weinberg equilibrium with most population-locus

SNP	Light	Blue	Inter light	Green-Hazel	Inter dark	Brown	Dark
rs12913832	1.63E-22	9.07E-08	0.000982285	1.34E-16	4.31E-08	1.02E-07	1.23E-19
rs1129038	1.90E-11	9.16E-08	0.001813796	1.77E-06	0.966208929	4.94E-09	4.68E-08
rs1667394	0.98656339	0.98897797	0.992225296	5.63E-05	0.033687951	0.022760921	0.010891357
rs916977	2.46E-07	0.00082391	0.992428401	4.92E-07	0.021157991	3.43E-13	0.000111655
rs4778138	2.21E-06	0.00318509	0.010009579	0.141407228	0.021614148	0.015505825	1.17E-06
rs7495174	0.000640462	0.98844347	0.100825555	0.801982186	3.90E-05	1.86E-09	0.016978639
rs4778241	1.34E-07	0.001065	0.006122944	0.07291421	1.20E-07	3.68E-08	0.005464988
rs4778232	0.045021572	0.02427662	0.041740677	0.242247974	0.082383243	8.86E-07	0.010570034
rs8024968	0.002116001	0.00026506	0.084192808	0.763037498	0.003484741	0.057402954	0.017411488
rs11636232	0.002232486	0.00984787	0.06079874	0.597900646	1.55E-05	1.08E-06	0.010891357
rs7183877	4.67E-06	0.00040978	0.431862256	0.020570431	0.000632146	0.002193017	1.68E-07
rs683	0.00785713	0.00479457	0.747794021	0.468583099	0.418201869	0.013356773	0.000115553
rs1375164	0.002381822	0.04988424	0.157804374	0.418678863	0.021377198	0.008423432	0.018563452
rs1800407	0.008872451	0.00962987	0.503888571	0.085648505	0.907231017	0.047848898	0.126259068
rs1408799	0.002538507	0.00145648	0.160767465	0.423365505	0.481717722	0.041197784	0.137183467
rs12896399	0.006109345	0.0010307	0.654440111	0.029891178	0.394995116	0.980154085	0.492793997
rs16891982	0.889890068	0.228141	0.024861456	0.021288447	0.988038577	0.007849112	0.059007875
rs26722	0.367022929	0.0417483	0.020695426	0.014288689	0.117339255	0.109819314	0.011320081
rs12203592	0.497122416	0.85151945	0.6288359	0.828822153	0.404212393	0.700206953	0.291660463
rs6058017	0.483001957	0.4252829	0.983345644	0.447677036	0.100148293	0.586630757	0.060142119
rs12592730	0.063724783	0.35652553	0.997039321	0.116832161	0.093355341	0.775592088	0.097462103
rs1393350	0.022375633	0.4402755	0.782477898	0.019342046	0.2993267	0.676743221	0.729194412
rs1015362	0.42197871	0.16975601	0.636160655	0.052506318	0.526398902	0.005100181	0.045148129

High

Low

Additional effect adjusting for rs12913832

Associated (p<0.05)

No effect (p>0.05)

Fig. 2. A schematic representation of the individual association analysis of 23 SNP components for each eye color class (orange cells) and their association after adjustment with rs12913832 (dark-orange cells), SNPs are listed in order of descending divergence. Rows represent each color vs. all other eye colors.

combinations not indicating significant departure from Hardy–Weinberg equilibrium and only six significant values but distributed randomly across eye color ‘populations.’

From 2530 SNP pair comparisons significant LD was found in 40 SNP pairs listed in [supplementary Table S2](#). All cases of significant LD were randomly distributed across loci and eye color populations except SNPs rs1375164–rs4778232 in OCA2 (15:28291812–15:28281765 = 10,047 nucleotide separation).

The PCA analysis did not reveal discernable geographic stratification – i.e., the phenotypic classes that we assigned to samples repeatedly cluster independently of their European geographic distribution as indicated in [Fig. 1C and D](#).

Handling the phenotypes observed as genetic populations and using blue and brown classes as pre-defined populations, *Structure* analysis suggested a pattern of continuity observed in intermediate phenotypes distinct from blue and brown, represented in the K:2 cluster plot of [Fig. 1A](#). We have positioned iris photographs above each sample range to illustrate the characteristics of the phenotypes assigned in the study. The intermediate-light samples tended to cluster close to the blue end and the intermediate-dark samples mostly clustered with browns; green-hazels appear as a largely equal mixture of the two most differentiated classes of blue and brown. This suggests that the simplified light and dark eye color phenotyping regime of Menger-From continues to have validity for the wider range of SNPs examined in our study. Nevertheless, the green-hazel phenotype does not form a distinct cluster at K:2, K:3, or K:4 (K:3 and 4 cluster plots in [supplementary Fig. S3](#)) and appears as a group where all components have mixed membership proportions. Furthermore, the optimum K estimation indicated that more than two genetic populations were detectable (probability plot in [Fig. 1B](#)). Therefore we adopted the green-hazel class as a third reference phenotype in the training sets. In order to

ensure the clearest differentiation of samples among a continuum of very similar iris colors, we excluded 35 green-hazel samples whose iris images fell between the most clearly differentiated blue and green-hazel phenotypes. Similarly, a further 13 brown eye samples were excluded that fell between green-hazel and brown phenotypes. So the training sets (256 samples) comprised: 145 blue subjects; 60 green-hazel from an original 95; 51 brown from an original 64.

3.2. Association analysis

The *p*-values for association with light, dark and the five eye color phenotypes originally defined in our sample set, for all 37 SNPs of SHEP 1–2, are given in [supplementary Table S3](#) and this data includes adjustment for the effect of rs12913832. Of the 14 additional skin and hair color-associated SNPs genotyped in the SHEP assays, 6 of 98 association *p*-values were significant ($p < 0.05$) after adjustment for rs12913832, but we did not pursue the analysis of these SNPs further. The levels of association found in the 23 SNPs previously identified as most closely associated with eye color are summarized in [Fig. 2](#). Association with a significance value of $p < 0.05$ was recorded in twenty markers. The independent additional effect on the prediction of phenotypes was measured by adjustment with the most associated SNP rs12913832. From adjustment analysis SNPs were detected to give an additional effect differentiating light, dark and the five eye color phenotypes independently of their linkage to rs12913832, all located in the OCA2/HERC2 region – in ranked order of effect: rs1129038, rs1667394, rs916977, rs4778138, rs7495174, rs4778241, rs4778232, rs8024968, rs11636232, rs7183877. These were followed by associated SNPs in genes: SLC45A2 (rs26722, rs16891982), SLC24A4 (rs12896399), ASIP (rs1015362), TYR

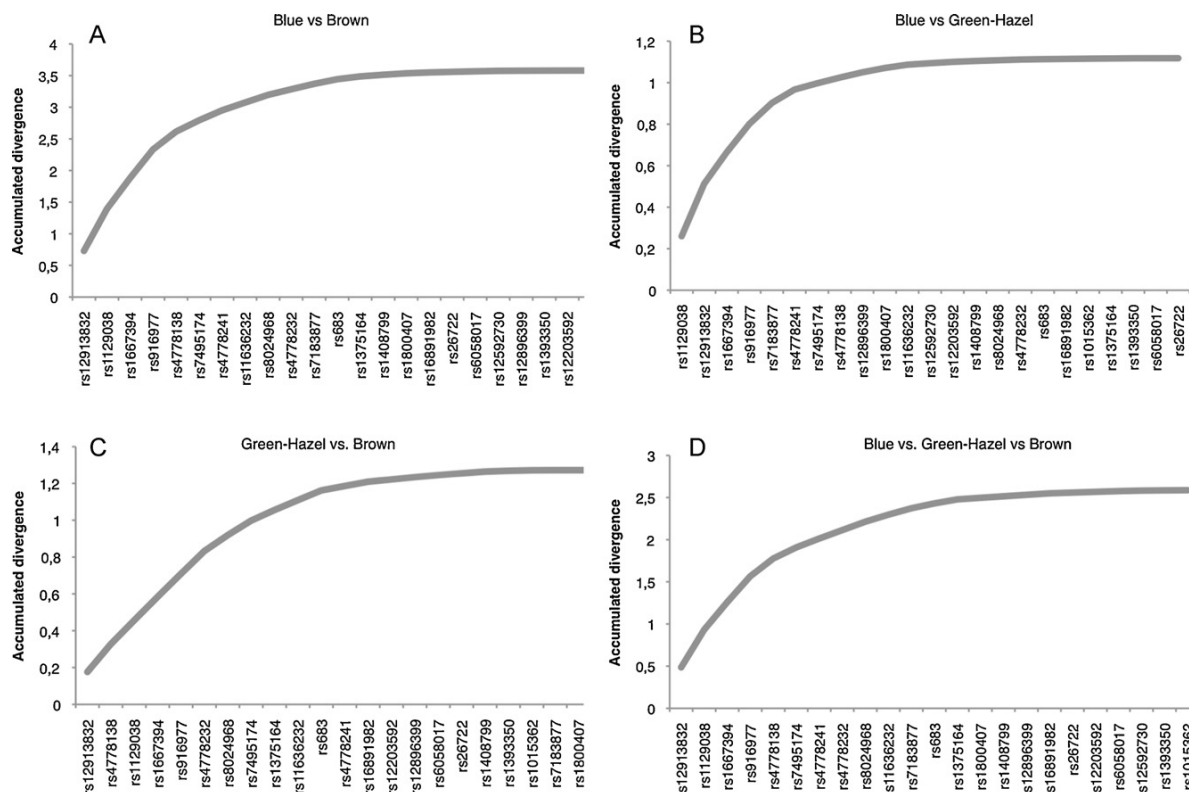


Fig. 3. Accumulated divergence values from classifications using the Bayesian system for pairwise eye color comparisons (A–C) and three-way eye color comparisons (D).

(rs1393350), and TYRP1 (rs1408799). Comparing eye color phenotyping approaches, the classes light and dark have the largest number of SNPs associated after adjustment for rs12913832. Additional adjusted association values for the other eleven SNPs most strongly associated in the above data were obtained, by treating the rs12913832–rs1129038 SNP pair as a single variant and as two variants, and *p*-values are given in [supplementary Table S4](#). Results were consistent between both approaches indicating strongest adjusted associations were for light and dark phenotypes with rs1667394 the strongest adjusted association to dark eye color.

3.3. Comparative assessment of classification with different SNP predictors

The divergence between the three phenotypes of the training set for 23 markers, tested with Jensen and Shannon's divergence metric, provided a ranking of SNP differentiation showing rs12913832 to be by far the most divergent, in agreement with a large number of previous studies [13,14,17–22,29,30]. SNP rs12913832 is followed by rs1129038, rs1667394, rs916977, rs4778138, rs749517, rs4778241, and rs4778232, all located in the region of the HERC2–OCA2 complex. This divergence data provides the order of the SNPs in the three plots of [Fig. 3](#), indicating the relative predictive power of each of the 23 SNPs from left (strongest) to right. The cumulative divergence values were highest for the differentiation of blue and brown compared to green-hazel with blue, or green-hazel with brown, but it is also notable that divergence values differed considerably between

component SNPs. HERC2 loci rs12913832 and rs1129038 gave divergences of ~ 0.7 for blue-brown comparisons. Although rs1667394 gave consistently high values, over half the other SNPs in any one comparison had much lower divergence values of 0.01–0.002 and rs6058017 of ASIP was the least informative marker. The allele frequencies observed in the eye color populations are listed in [supplementary Table S3](#).

Evaluating ROC curves as an alternative phenotype assignment method to *Snipper* allowed a direct comparison with the system used by Walsh et al. based on the six *Irisplex* SNPs, all included in our set. AUC estimations of the ROC curves were made for each pairwise phenotype comparison and these are shown in [Fig. 4](#) for varying SNP subsets. In agreement with the findings of Walsh et al., the predictions for blue and brown eye colors were mostly explained by rs12913832 producing AUC values of 0.952 and 0.965 respectively, from this single marker alone. For green-hazel predictions additional SNPs are required to improve the precision but these do not exceed AUC values above 76% with six or indeed eight SNPs, as shown in [Fig. 4A](#) and B. We reach similar AUC values to those of Walsh with values of 0.986 for blue, 0.756 for green-hazel, and 0.978 for brown ([Fig. 4A](#)). However aiming to achieve a balance between small numbers of markers and predictive value we found that the addition of other markers, shown to be independent in effect from rs12913832, raised the AUC values for all phenotypes enough to justify their inclusion. This was particularly noticeable moving from the two additional SNPs of Mengel-From to a total of 13 that included SNPs improving green-hazel prediction. Adding the two additional SNPs of Mengel-From gave AUC values of 0.996 for blue and 0.983 for brown, plus slight

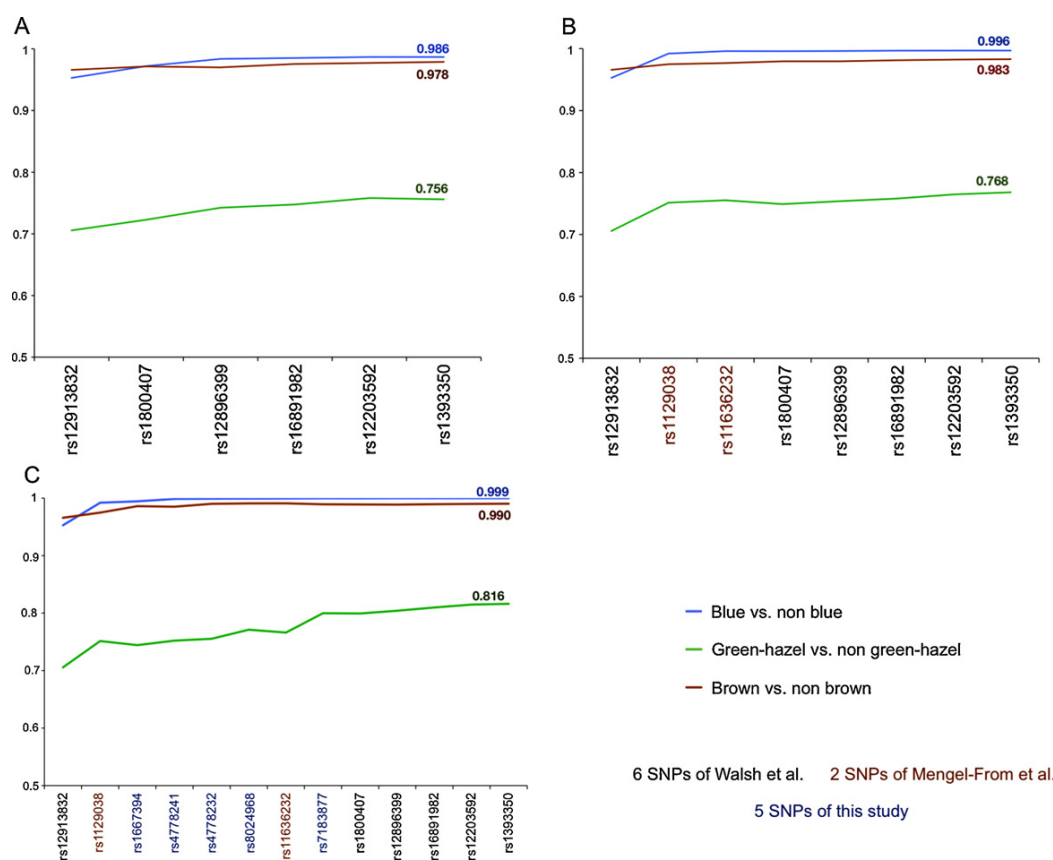


Fig. 4. AUC curves for: (A) 6 SNP predictors of Walsh et al.; (B) 6 + 2 SNP predictors of Mengel-From et al.; (C) 6 + 2 plus the five additional SNP predictors identified in this study.

progression to 0.768 for green-hazel (Fig. 4B). When the additional five SNPs were incorporated to the analysis, predictability from the AUC value estimates increases significantly to: 0.999 for blue, 0.990 for brown and 0.816 for green hazel (Fig. 4C). In summary, there is evidence the incorporation of rs1129038 of HERC2 improves eye color predictability in general, while adding rs1667394 and rs7183877 (HERC2) improves hazel-green predictability.

3.4. MDR analysis

MDR analysis indicated that when assessing all possible combinations of two, three, and four SNPs the best model observed was composed of two SNPs for each eye color comparison (i.e., brown vs. non-brown, etc.). In brown vs. non-brown comparisons, the best model included rs12913832 in HERC2 and rs4778138 in OCA2 (BA 0.8587, CV 10/10). The entropy analysis indicated an interaction effect between these SNPs that was redundant as shown by the blue lines of the right-hand dendrogram of Fig. 5. For green-hazel vs. non green-hazel comparisons, the SNPs selected in the best model were HERC2 rs12913832 and rs1667394 (BA 0.8322, CV 8/10) with a significance level of $p < 0.001$. In this case a synergistic interaction was observed between these SNPs indicated by the red lines on the left-hand dendrogram of Fig. 5B. Results were confirmed when these selected SNPs and their interactions were included in a logistic regression under additive model, with a significance p -value of 2.03×10^{-15} for green-hazel and 0.00130 for brown (data not shown).

Curiously the interactive effect between rs12913832 and rs1667394 is evidently different between the two eye color comparisons shown in Fig. 5. Such complex interactive effects require detailed further analysis and we are currently pursuing these studies.

3.5. Predictive performance

The informativeness of the reference samples used as a 23-SNP training set for the *Snipper* classifier was tested by cross-validation and gave classification success rates of: blue 97.93%; green-hazel 98.93%; and brown 92.16%. The modified bootstrap analysis revealed comparable success rates of: blue 98.27%, green-hazel 97.81%, and brown 96.67%. These measurements of classification success/error support the approach used to create three distinct reference classes for the classification system, despite excluding a proportion of complex intermediate phenotypes from the reference set.

The results based on the AUC curves, outlined in Section 3.3 above, indicated rs1129038 improves the predictive performance of the six *Irisplex* SNPs. There was also a marked jump in AUC value for green-hazel with the inclusion of rs7183877. Additionally, the MDR results in Section 3.4 suggest a synergistic interaction between rs12913832 and rs1664394. Therefore we examined the effect on sensitivity, specificity, NPV and PPV, for three SNP sets: *Irisplex* 6; *Irisplex* 5 with the rs12913832–rs1129038 pair and; *Irisplex* 5 + 2 with HERC2 SNPs rs7183877 and rs1664394.

Table 3 details these measures of predictive performance obtained by making *Snipper* classifications with the three SNP sets. The table indicates the effect of adding rs1129038 or three additional HERC2 SNPs is to improve the sensitivity of green-hazel predictions, that jump from 3.3% (Table 3A) to 75.3% (Table 3B) with the addition of rs1129038. Specificity is also improved for light, blue, intermediate-light, intermediate-dark, brown, and dark (Table 3B). Overall, each specificity is raised in value when adding rs1129038 except green-hazel, however this phenotype gains the most improved balance between sensitivity and specificity. Table 3C indicates the addition of further HERC2 SNPs does not necessarily improve the overall classification performance of

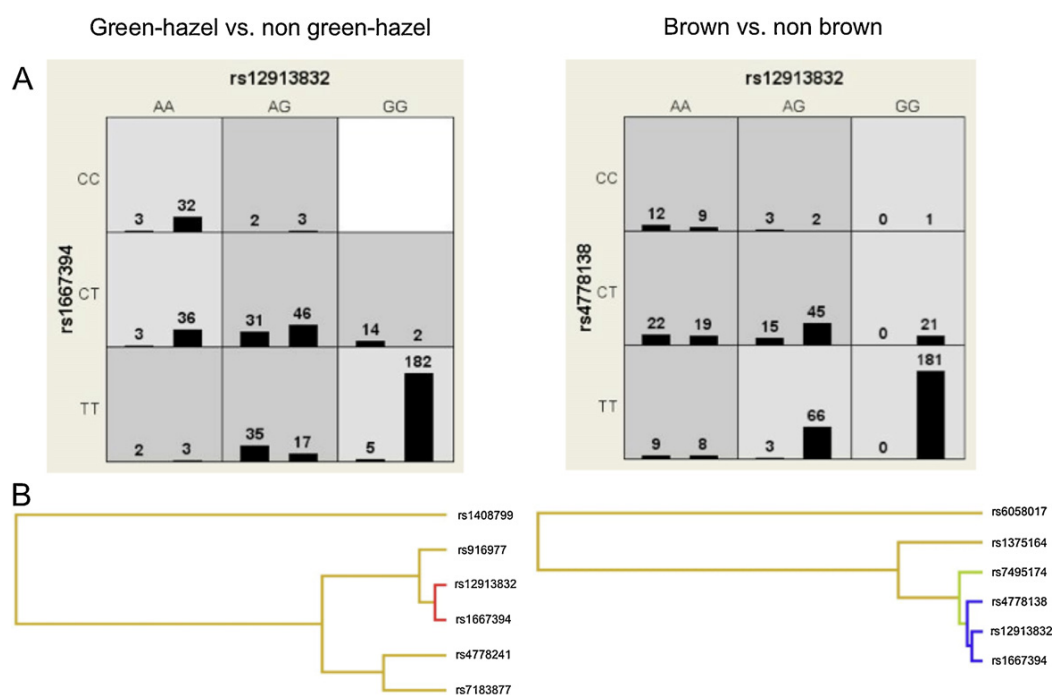


Fig. 5. Analysis of SNP interactions with MDR: (A) interaction of SNP-pair combinations and (B) corresponding entropic dendrograms for green-hazel vs. non green-hazel and brown vs. non-brown comparisons. In the interaction charts left bars indicate 'cases' (i.e., eye color) and right bars indicate the 'controls' (i.e., all other eye colors), light cells represent low risk and dark cells high risk. In the left-hand dendrogram the red lines represent synergistic interactions and on the right, blue lines represent redundant interactions.

Table 3

AUC % predictive value estimates from *Snipper* classifications for eye colors from pairwise comparisons of the full range of eye colors with three sets of SNP predictors: (A) *Irisplex* 6; (B) adding rs12913832–rs1129038 as a SNP pair PPV: positive predictive value, NPV: negative predicted value, ND: not determined. Bold values indicate the perceived best balance between sensitivity and specificity.

	Light	Blue	Inter-light	Green-hazel	Inter-dark	Brown	Dark
(A) 6 <i>Irisplex</i> SNPs							
Sensitivity	99.41	99.26	100.00	3.33	82.50	97.67	90.36
Specificity	85.84	66.22	39.52	98.02	75.62	79.08	91.46
PPV	91.30	72.83	18.48	16.67	35.87	45.65	81.52
NPV	98.98	98.99	100.00	89.49	96.32	99.47	95.79
(B) <i>Irisplex</i> with rs12913832–rs1129038							
Sensitivity	96.51	98.53	88.89	75.32	50.82	81.67	66.12
Specificity	98.99	85.59	59.28	83.96	78.64	84.52	93.17
PPV	98.81	79.76	19.05	55.24	31.96	50.52	82.47
NPV	97.03	99.02	98.02	92.83	89.01	95.97	84.98
(C) <i>Irisplex</i> with 4 <i>HERC2</i> SNPs							
Sensitivity	95.00	98.58	82.50	13.04	75.00	89.09	81.10
Specificity	78.06	68.07	45.99	94.16	73.44	73.29	87.20
PPV	79.91	64.65	15.35	33.33	40.00	36.30	76.30
NPV	94.44	98.78	95.68	82.86	92.56	97.52	90.08

Snipper and does not gain better balance between sensitivity and specificity. Since predictive models do not take account of interactive effects this may explain the lack of improvement in the sensitivity/specificity values when including two additional *HERC2* SNPs (the four in total of Table 3C).

4. Discussion

In the present study of six European populations, we confirmed the effect of most of the SNP predictors previously published for blue/brown or light/dark iris colors [14,19]. Specifically rs12913832, rs1129038, and rs116363232 in *HERC2* and rs1289399 in *SLC24A4* showed strong associations when making appropriate adjustment for the effect of rs12913832 or rs12913832–rs1129038. The markers rs1800407 in *OCA2*, rs16891982 in *SLC45A2*, and rs1393350 in *TYR* are also good predictors for forensic tests. The SNP rs12203592 in *IRF4*, reported to be among the six best predictors by Walsh et al., was not found to be a good predictor in our study. Interestingly, the frequency of the minor allele of rs12203592 had a high frequency in the Northern Ireland sample analyzed recently in an extended study of *Irisplex* in European populations [44] but has a much lower frequency in the samples of our study. This may explain the weaker predictability of rs12203592 in our analyses and raises the issue of population differences within Europe that will require extended examination to properly gauge the value of SNPs outside of the *OCA2*–*HERC2* complex. We also confirmed the previously reported AUC values for the six *Irisplex* SNPs of Walsh proposed for forensic blue/brown eye color prediction. When considering the extension of these six predictors our study findings lead us to recommend the inclusion of *HERC2* SNP rs1129038. This SNP improves the predictability for all eye colors as indicated by the AUC analyses and we have shown it has an additional effect to rs12913832 after adjustment analysis. In addition to rs1129038 the predictability of complex intermediate phenotypes increased when incorporating markers, in order of decreasing divergence: rs916977, rs7183877 (*HERC2*), rs4778138, rs7495174, rs4778241, rs8024968, rs1375164 (*OCA2*), rs1408799 (*TYRP1*), and rs26722 (*SLC45A2*). In particular, rs7183877 makes a detectable contribution to the prediction of green-hazel eye color. We have also begun epistasis studies with initial MDR results reported here and these already point to a synergistic interactive effect within *HERC2* for rs1667394 with rs12913832 for green-hazel prediction. Therefore the three SNPs: rs1129038, rs1667394, and rs7183877, despite their close proximity in *HERC2*, provide additional information for eye color prediction after adjustment with rs12913832. This is

consistent with the idea of complex regulatory function interactions within the *OCA2*–*HERC2* complex [11,16,18,19], while the latter two SNPs were the 9th and 10th best predictors in the Liu study [19]. We note that including S European population samples in our study provided 50% non-blue, non-brown subjects, but these comprised just 10% of Liu's study. Incorporating rs1129038, rs7183877, and rs1667394 with the six SNPs of *Irisplex* had the most marked effect on green-hazel predictability, but we also achieved a slight improvement in the prediction of blue and brown eye colors (AUC levels of 99% for blue and 98% for brown). However, further studies analyzing the independence of the additional *OCA2*–*HERC2* markers closely located in the same chromosomal region are recommended, in particular utilizing the extra power provided by family studies.

The proper design and construction of a forensic eye color predictive test requires the consideration of certain factors examined in this study, in addition to exploring the effect of expanding the number of SNP predictors. Firstly, it is important to adequately define the phenotypic classes. We collected samples from six populations from north and south Europe with a range of phenotypic variability likely to be wider than would be found within single geographic locations. Reducing the complexity of the intermediate eye color range in the reference samples provided an informative training set and helped to identify key additional predictors. Nevertheless we found the AUC values for intermediate eye colors reached a plateau at ~82% with 13 SNPs. This indicates additional factors have yet to be identified for intermediate eye color expression, including: undetected SNPs and associations; epistatic interactions; additional pigmentation genes; and environmental effects (e.g., changing iris color with age, as suggested by a recent study which also explored several complex SNP–SNP interactions [7]). It is worth noting that *OCA2* is among the larger human genes (344 kb) and is likely to harbor many low frequency coding SNPs within the total 24 exons. Therefore, at the moment, a forensic eye color test maintaining a reasonably manageable multiplex level has probably reached the limit of predictive power for intermediate eye colors. The problem of a uniform, objective recognition of intermediate iris colors distinct from brown/blue among a group of observers (such as eye witnesses) also remains a source of variation in predictive performance of forensic tests, over and above complex SNP associations. Secondly, of equal importance to the definition of phenotypes is the classification method used. For routine inference of eye color from a forensic test the classifier must have a degree of flexibility that allows the end-user to re-configure SNP profiles according to their operational needs, as well as the scope to include extra markers from newly

discovered or closely linked SNP associations. Luckily, both the classifiers of Walsh et al. and the online Bayesian system proposed here allow the user to make their own decisions about a probability threshold: a value limit that balances the 'benefit' of sufficiently successful predictions against the 'cost' of too many erroneous predictions or too many that are undefined because of an unduly high probability cut-off. These limits correspond to suggested values of 0.7 using the Walsh classifier and a likelihood ratio of 3:1 we applied using *Snipper*. However it may be the case that users wish to explore the effect of extra SNPs using their own test sets and *Snipper* allows likelihoods to be collected for such test sets analyzing a range of marker combinations. Though we used AUC analysis in order to compare the performance of extra OCA2–HERC2 SNPs with the established six markers of Walsh, we actually found *Snipper* easier to use to assess each new marker one at a time, but more importantly to upload linked SNP data by applying the SNP rs12913832–rs1129038 pair frequencies observed in each training set phenotype. Another advantage of *Snipper* is the ability to deal with incomplete profiles that have missing SNP data, common when analyzing highly degraded DNA. Therefore profiles where weakly predictive SNPs are missing are valid and likely to obtain high probabilities for blue or brown eye colors.

Acknowledgments

YR was supported by the Fundación Gran Mariscal de Ayacucho (FUNDAYACUCHO). MVL was supported by funding from Xunta de Galicia INCITE 09 208163PR and this work was in part supported by additional funding from Xunta de Galicia: PGIDIT06P-XIB228195PR. JS was supported by the German Academic Exchange Service (DAAD). AC and CP acknowledge the support of the Areces Foundation and the EuroForGen NoE. The authors would like to thank all the anonymous donors who participated in the study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.fsigen.2012.05.009>.

References

- [1] A. Heaton-Armstrong, Eye-witness testimony and judicial studies, *Med. Sci. Law* 35 (1995) 93–94.
- [2] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, *Forensic Sci. Int. Genet.* 3 (2009) 154–161.
- [3] R.A. Sturm, T.N. Frudakis, Eye color: portals into pigmentation genes and ancestry, *Trends Genet.* 20 (2004) 327–332.
- [4] D.M. Albert, W.R. Green, M.L. Zimbric, C. Lo, R.E. Gangnon, K.L. Hope, J. Gleiser, Iris melanocyte numbers in Asian, African American, and Caucasian irides, *Trans. Am. Ophthalmol. Soc.* 101 (2003) 217–222.
- [5] P.D. Imesch, I.H. Wallow, D.M. Albert, The color of the human eye: a review of morphologic correlates and of some conditions that affect iridial pigmentation, *Surv. Ophthalmol.* 41 (Suppl. 2) (1997) S117–S123.
- [6] R.A. Sturm, M. Larsson, Genetics of human iris color and patterns, *Pigment Cell Melanoma Res.* 22 (2009) 544–562.
- [7] F. Liu, A. Wollstein, P.G. Hysi, G.A. Ankrá-Badu, T.D. Spector, D. Park, G. Zhu, M. Larsson, D.L. Duffy, G.W. Montgomery, D.A. Mackey, S. Walsh, O. Lao, A. Hofman, F. Rivadeneira, J.R. Vingerling, A.G. Uitterlinden, N.G. Martin, C.J. Hammond, M. Kayser, Digital quantification of human eye color highlights genetic association of three new loci, *PLoS Genet.* 6 (2010) e1000934.
- [8] R.K. Valenzuela, M.S. Henderson, M.H. Walsh, N.A. Garrison, J.T. Kelch, O. Cohen-Barak, D.T. Erickson, F. John Meaney, J.B. Walsh, K.C. Cheng, S. Ito, K. Wakamatsu, T. Frudakis, M. Thomas, M.H. Brilliant, Predicting phenotype from genotype: normal pigmentation, *J. Forensic Sci.* 55 (2010) 315–322.
- [9] T. Frudakis, T. Terravainen, M. Thomas, Multilocus OCA2 genotypes specify human iris colors, *Hum. Genet.* 122 (2007) 311–326.
- [10] T. Frudakis, M. Thomas, Z. Gaskin, K. Venkateswarlu, K.S. Chandra, S. Ginjupalli, S. Gunturi, S. Natrajan, V.K. Ponnuswamy, K.N. Ponnuswamy, Sequences associated with human iris pigmentation, *Genetics* 165 (2003) 2071–2083.
- [11] D.L. Duffy, N.F. Box, W. Chen, J.S. Palmer, G.W. Montgomery, M.R. James, N.K. Hayward, N.G. Martin, R.A. Sturm, Interactive effects of MC1R and OCA2 on melanoma risk phenotypes, *Hum. Mol. Genet.* 13 (2004) 447–461.
- [12] D.L. Duffy, G.W. Montgomery, W. Chen, Z.Z. Zhao, L. Le, M.R. James, N.K. Hayward, N.G. Martin, R.A. Sturm, A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation, *Am. J. Hum. Genet.* 80 (2007) 241–252.
- [13] R.A. Sturm, D.L. Duffy, Z.Z. Zhao, F.P. Leite, M.S. Stark, N.K. Hayward, N.G. Martin, G.W. Montgomery, A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color, *Am. J. Hum. Genet.* 82 (2008) 424–431.
- [14] H. Eiberg, J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K.W. Kjaer, L. Hansen, Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression, *Hum. Genet.* 123 (2008) 177–187.
- [15] M. Kayser, F. Liu, A.C. Janssens, F. Rivadeneira, O. Lao, K. van Duijn, M. Vermeulen, P. Arp, M.M. Jhamai, W.F. van Ijcken, J.T. den Dunnen, S. Heath, D. Zelenika, D.D. Despriet, C.C. Klaver, J.R. Vingerling, P.T. de Jong, A. Hofman, Y.S. Aulchenko, A.G. Uitterlinden, B.A. Oostra, C.M. van Duijn, Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene, *Am. J. Hum. Genet.* 82 (2008) 411–423.
- [16] R.A. Sturm, Molecular genetics of human pigmentation diversity, *Hum. Mol. Genet.* 18 (2009) R9–R17.
- [17] J. Mengel-From, T.H. Wong, N. Morling, J.L. Rees, I.J. Jackson, Genetic determinants of hair and eye colors in the Scottish and Danish populations, *BMC Genet.* 10 (2009) 88.
- [18] J. Mengel-From, C. Borsting, J.J. Sanchez, H. Eiberg, N. Morling, Human eye color and HERC2, OCA2 and MTP, *Forensic Sci. Int. Genet.* 4 (2010) 323–328.
- [19] F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, A.C. Janssens, M. Kayser, Eye color and the prediction of complex phenotypes from genotypes, *Curr. Biol.* 19 (2009) R192–R193.
- [20] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye color in the absence of ancestry information, *Forensic Sci. Int. Genet.* 5 (2010) 170–180.
- [21] W. Branicki, U. Brudnik, A. Wojas-Pelc, Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype, *Ann. Hum. Genet.* 73 (2009) 160–170.
- [22] E. Pospiech, J. Draus-Barini, T. Kupiec, A. Wojas-Pelc, W. Branicki, Gene–gene interactions contribute to eye color variation in humans, *J. Hum. Genet.* 56 (2011) 447–455.
- [23] C. Phillips, A. Salas, J.J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, M. Calaza, M.C. de Cal, D. Ballard, M.V. Lareu, A. Carracedo, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, *Forensic Sci. Int. Genet.* 1 (2007) 273–280.
- [24] N. Eriksson, J.M. Macpherson, J.Y. Tung, L.S. Hon, B. Naughton, S. Saxonov, L. Avey, A. Wojcicki, I. Pe'er, J. Mountain, Web-based, participant-driven studies yield novel genetic associations for common traits, *PLoS Genet.* 6 (2010) e1000993.
- [25] O. Lao, J.M. de Gruijter, K. van Duijn, A. Navarro, M. Kayser, Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms, *Ann. Hum. Genet.* 71 (2007) 354–369.
- [26] H.L. Norton, R.A. Kittles, E. Parra, P. McKeigue, X. Mao, K. Cheng, V.A. Canfield, D.G. Bradley, B. McEvoy, M.D. Shriver, Genetic evidence for the convergent evolution of light skin in Europeans and East Asians, *Mol. Biol. Evol.* 24 (2007) 710–722.
- [27] J. Graf, J. Voisey, I. Hughes, A. van Daal, Promoter polymorphisms in the MTP (SLC45A2) gene are associated with normal human skin color variation, *Hum. Mutat.* 28 (2007) 710–717.
- [28] J. Graf, R. Hodgson, A. van Daal, Single nucleotide polymorphisms in the MTP gene are associated with normal human pigmentation variation, *Hum. Mutat.* 25 (2005) 278–284.
- [29] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, K.P. Magnusson, A. Manolescu, A. Karason, A. Palsson, G. Thorleifsson, M. Jakobsdottir, S. Steinberg, S. Palsson, F. Jonasson, B. Sigurgeirsson, K. Thorisdottir, R. Ragnarsson, K.R. Benediktsson, K.K. Aben, L.A. Kiemeny, J.H. Olafsson, J. Gulcher, A. Kong, U. Thorsteinsdottir, K. Stefansson, Genetic determinants of hair, eye and skin pigmentation in Europeans, *Nat. Genet.* 39 (2007) 1443–1452.
- [30] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, M. Jakobsdottir, S. Steinberg, S.A. Gudjonsson, A. Palsson, G. Thorleifsson, S. Palsson, B. Sigurgeirsson, K. Thorisdottir, R. Ragnarsson, K.R. Benediktsson, K.K. Aben, S.H. Vermeulen, A.M. Goldstein, M.A. Tucker, L.A. Kiemeny, J.H. Olafsson, J. Gulcher, A. Kong, U. Thorsteinsdottir, K. Stefansson, Two newly identified genetic determinants of pigmentation in Europeans, *Nat. Genet.* 40 (2008) 835–837.
- [31] R.P. Stokowski, P.V. Pant, T. Dadd, A. Fereday, D.A. Hinds, C. Jarman, W. Filsell, R.S. Ginger, M.R. Green, F.J. van der Ouderaa, D.R. Cox, A genomewide association study of skin pigmentation in a South Asian population, *Am. J. Hum. Genet.* 81 (2007) 1119–1132.
- [32] J. Han, P. Kraft, H. Nan, Q. Guo, C. Chen, A. Qureshi, S.E. Hankinson, F.B. Hu, D.L. Duffy, Z.Z. Zhao, N.G. Martin, G.W. Montgomery, N.K. Hayward, G. Thomas, R.N. Hoover, S. Chanock, D.J. Hunter, A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation, *PLoS Genet.* 4 (2008) e1000074.
- [33] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, *Methods Mol. Biol.* 132 (2000) 365–386.
- [34] P.M. Vallone, J.M. Butler, AutoDimer: a screening tool for primer–dimer and hairpin structures, *Biotechniques* 37 (2004) 226–231.

- [35] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, *Genetics* 155 (2000) 945–959.
- [36] N.A. Rosenberg, DISTRUCT: a program for the graphical display of population structure, *Mol. Ecol. Notes* 4 (2004) 137–138.
- [37] F. Curtin, P. Schulz, Multiple correlations and Bonferroni's correction, *Biol. Psychiatry* 44 (1998) 775–777.
- [38] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver, H.E. Stanley, Analysis of symbolic sequences using the Jensen–Shannon divergence, *Phys. Rev. E: Stat. Nonlin. Soft Matter Phys.* 65 (2002) 041905.
- [39] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, *Am. J. Hum. Genet.* 73 (2003) 1402–1422.
- [40] C.F.J. Wu, The Jackknife, the bootstrap and other resampling methods in regression analysis, *Ann. Stat.* 14 (1982) 1261–1295.
- [41] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCr: visualizing classifier performance in R, *Bioinformatics* 21 (2005) 3940–3941.
- [42] J.H. Moore, J.C. Gilbert, C.T. Tsai, F.T. Chiang, T. Holden, N. Barney, B.C. White, A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, *J. Theor. Biol.* 241 (2006) 252–261.
- [43] P. Frost, European hair and eye color: a case of frequency-dependent sexual selection? *Evol. Hum. Behav.* 27 (2006) 85–103.
- [44] S. Walsh, A. Wollstein, F. Liu, U. Chakravarthy, M. Rahu, J.H. Seland, G. Soubrane, L. Tomazzoli, F. Topouzis, J.R. Vingerling, J. Vioque, A.E. Fletcher, K.N. Ballantyne, M. Kayser, DNA-based eye colour prediction across Europe with the IrisPlex system, *Forensic Sci. Int. Genet.* 6 (2012) 330–340.

Bloque 2.

V.6. A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type

O. Bulbul, G. Filoglu, H. Altuncul, A. FreireAradas, Y. Ruiz, M. Fondevila, C. Phillips, Á.Carracedo, A.K. Kriegel, P.M. Schneider

(Forensic Science International:GeneticsSupplementSeries, 2011, 3 e500–e501)



Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/FSIGSS

A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type

O. Bulbul^{a,c,*}, G. Filoglu^a, H. Altuncul^a, A. Freire Aradas^b, Y. Ruiz^b, M. Fondevila^b, C. Phillips^b, Á. Carracedo^b, A.K. Kriegel^c, P.M. Schneider^c

^a Institute of Forensic Sciences, Istanbul University, Turkey

^b Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Spain

^c Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Germany

ARTICLE INFO

Article history:

Received 27 September 2011

Accepted 2 October 2011

Keywords:

Ancestry-informative marker

AIMs

Pigmentation

SNPs

Eye colour

ABSTRACT

DNA analysis of ancestry informative markers (AIMs) and physical trait markers from biological stains can help provide investigative leads in cases without suspects. To enhance the resolution and informative value of two previously developed single nucleotide polymorphism (SNP) multiplexes, the 34-plex and Eurasiaplex assays (Phillips et al. [1], Phillips et al. [2]) differentiating European, South Asian and Middle East populations, we have selected an additional 22 AIM-SNPs. The selected markers focus on the differentiation of Europeans and Asians and we supplemented this AIM set with 10 recently published pigmentation markers informative for eye, hair and skin colour (Walsh et al. [4], Sturm [3]). Comparisons of reference SNP data from HGDP-CEPH and 1000 Genomes population panels with 7 Eurasian study populations were made using STRUCTURE and principal component analysis (PCA) indicating that this multiplex can improve the differentiation of populations within East Asia. In a small pilot study using voluntary donors from Turkey, the IrisPlex SNP markers (Walsh et al. [4]) built into our multiplex were successfully used to make predictions about eye colour.

© 2011 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The aim of this study was to differentiate the ancestry of closely related Eurasian subpopulations using small-scale forensic AIM-SNP sets complementary to two other multiplexes, a 34-plex AIM set in established use [1] and a 23-plex: Eurasiaplex in development [2]. The 34-plex set was primarily designed to distinguish African, European and East Asian ancestries. Two new sets: Eurasiaplex and the 32-plex multiplex presented here, have been developed to complement 34-plex to provide more clearly differentiated ancestry analyses of NW European, SE European, Middle Eastern and Central South Asian regions of Eurasia. We also successfully incorporated into the 32-plex assay ten key pigmentation predictive SNPs for eye and hair colour plus skin tone variation within Eurasian populations.

2. Materials and methods

From scrutiny of HapMap and dbSNP data we selected 22 new ancestry informative marker showing marked allele frequency differences between Europeans and East Asians with the aim of improving the differentiation of Europeans from Middle East and South East Asian populations sited in the middle of Europe and East Asia. The ten SNPs informative for eye, hair and skin colour phenotypes were selected directly from the current literature [3,4].

PCR primers were designed to give amplicon sizes from 75 to 148 bp. Cycling was performed in a Gene Amp 9700 thermal cycler with steps: denaturation at 95 °C for 10 min then 35 cycles of 95 °C for 30 s, 60 °C for 50 s and 65 °C for 40 s, then 6 min at 65 °C. Excess PCR primers and dNTPs were removed by adding ExoSAP-IT and incubating at 37 °C for 45 min then 85 °C for 15 min. Single base extensions (SBE) were performed using the SNaPshot Multiplex kit (Applied Biosystems) in 6 µl reaction volumes. SBE comprised 30 cycles of 96 °C for 10 s, 55 °C for 5 s, and 60 °C for 30 s. Excess nucleotides were removed with shrimp alkaline phosphatase (SAP) incubating at 37 °C for 80 min then 85 °C for 15 min. The SBE products were separated using an ABI 3130.

SNP genotypes were analysed with STRUCTURE software applied to 34-plex alone, 34-plex + Eurasiaplex, and 34-plex + Eurasiaplex + 32-plex (85 SNPs in total) of Turkish (56), Indian

* Corresponding author at: Institute of Forensic Science, Istanbul University, Istanbul Turkey.

E-mail address: oslembibl@yahoo.com (O. Bulbul).

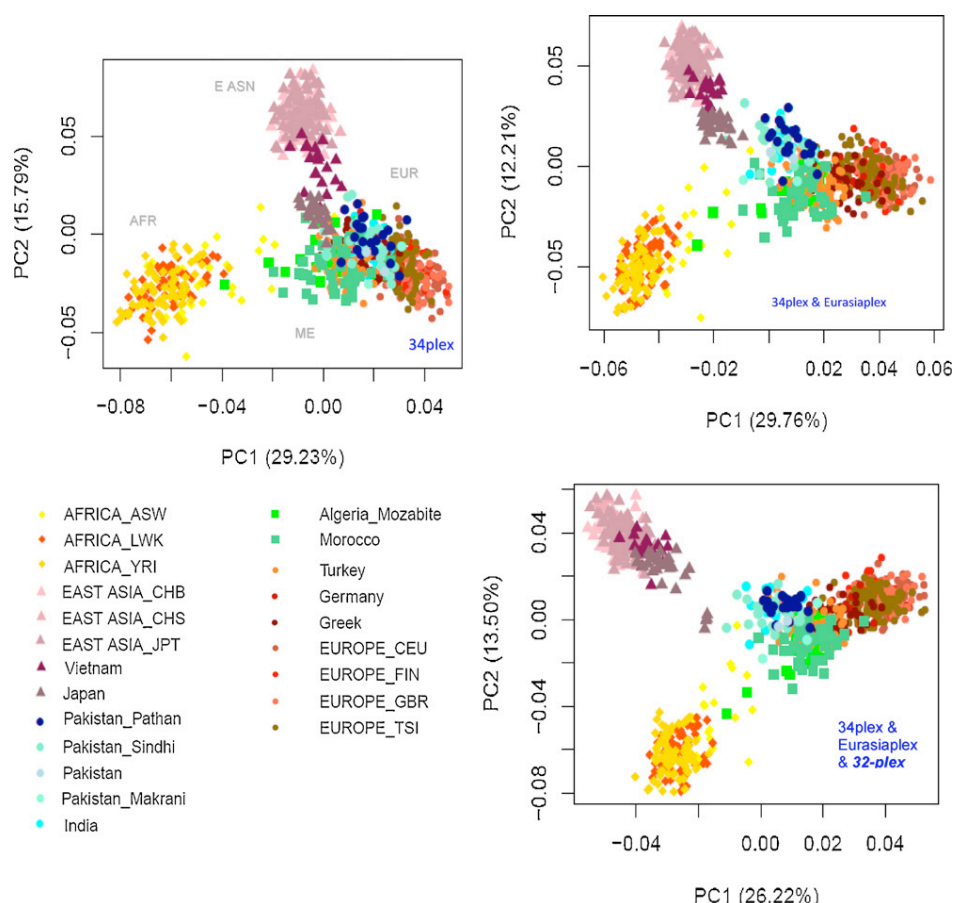


Fig. 1. PCA analyses of 22 populations using 34, 57 and 85 SNPs.

(23), Algerian (27), Moroccan (41), Vietnamese (20), Pakistani (52), Japanese (28), German (10), plus 1000 Genomes Phase I reference samples comprising 607 samples from 10 populations. Structure results were compared to those of Principal Component Analysis (PCA) to examine the degree of divergence amongst study and reference populations.

3. Results and discussion

Comparisons of HGDP-CEPH and 1000 Genomes reference populations with seven Eurasian study populations were made using STRUCTURE and PCA. Although STRUCTURE analysis indicated that the optimum *K* value was three and therefore Eurasian subpopulation differentiation is difficult to achieve, PCA (Fig. 1) shows differentiation of South Asian from most European populations with the exception of Turkish, where some overlap persists. The Middle East populations were particularly difficult to differentiate with considerable overlap between their positions and those of European and South Asian populations. However the best separation of the other four groups: European, South Asian, East Asian and African is obtained using all three SNP assays and since this only involves running three PCR this provides an optimum SNP based system for ancestry inference of Eurasians, East Asians and Africans. We intend to pursue the combination of the above 85 SNPs with STR and indel data to further differentiate Eurasian populations in forensic ancestry analysis.

The six *IrisPlex* SNPs we incorporated were used to predict eye colour according to Walsh et al. [4] in the Turkish samples. The *IrisPlex* SNPs gave consistent predictions for most brown and blue

eye colour subjects, but failed to adequately predict intermediate eye colour. Therefore we tested the same SNPs with the USC “Snipper” web portal (<http://mathgene.usc.es/snipper/>) to determine eye colour likelihoods using subjects with blue, intermediate and brown eye colour as reference samples. The *Snipper* web portal was found to predict intermediate eye colour with higher likelihoods compared to the original analytical approach. But intermediate eye colour is highly complex and some of the differences in predictive performance between each approach could arise from differences in assigning the intermediate iris phenotype.

Conflict of interest

None.

References

- [1] C. Phillips, A. Salas, J.J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, A. Carracedo, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, *Forensic Sci. Int. Genet.* 1 (2007) 273–280.
- [2] C. Phillips, A. Freire-Aradas, A.K. Kriegl, M. Fondevila, O. Bulbul, C. Santos, F. Surulla-Rech, A. Carracedo, P. Schneider, M.V. Lareu, *Eurasiaplex*: a forensic SNP assay for differentiating European and South Asian ancestries, *Forensic Sci. Int. Genet.* 2011, in preparation.
- [3] R.A. Sturm, Molecular genetics of human pigmentation diversity, *Hum. Mol. Genet.* 18 (1) (2009) R9–R17 (review issue).
- [4] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, *IrisPlex*: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, *Forensic Sci. Int. Genet.* 5 (2010) 170–180.

Bloque 2.

V.7. A researcher's guide to STRUCTURE software: applications, parameter settings and supporting software

Liliana Porras-Hurtado, Yarimar Ruiz, Carla Santos, Christopher Phillips, Maria Victoria Lareu and Ángel Carracedo.

(Review in preparation)

1 **A researcher's guide to *STRUCTURE* software: applications, parameter settings**
2 **and supporting software.**

3
4 *Liliana Porras-Hurtado^{1,2*}, Yarimar Ruiz^{2*}, Carla Santos^{2*}, Christopher Phillips^{2†},*
5 *Maria Victoria Lareu² and Ángel Carracedo^{2,3}*

6
7 *¹Universidad Tecnológica de Pereira – Colombia*

8 *²Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de*
9 *Compostela, Santiago de Compostela, Galicia.*

10 *³Genomics Medicine Group, CIBERER, University of Santiago de Compostela,*
11 *Santiago de Compostela, Galicia, Spain*

12
13 *Keywords: population structure; case-control association studies; stratification;*
14 *STRUCTURE; CLUMPP; distruct; STRAT*

15
16 ** contributed equally to the work.*

17 *† communicating author: Christopher Phillips, Forensic Genetics Unit, Institute of Legal*
18 *Medicine, University of Santiago de Compostela, Santiago de Compostela, Galicia.*
19 *c.phillips@mac.com*

20
21 *Abstract word count: 181*

22
23 *Manuscript word count: 4,820*
24

ABSTRACT

STRUCTURE is widely used population analysis software that allows researchers to assess patterns of genetic structure in a set of samples. It can identify subsets of the whole sample by detecting allele frequency differences within the data and can assign individuals to those sub-populations based on analysis of likelihoods. We present an up-to-date review of *STRUCTURE*'s most commonly used ancestry and frequency models, plus an overview of the main applications of the software including case-control association studies, population genetics, forensic analysis and non-human population research. A detailed step-by-step guide to running *STRUCTURE* is provided in the accompanying supplementary file. With reference to a worked example, the guide explores the effects of changing the principal analysis parameters on *STRUCTURE* results. Additionally, descriptions and guidelines for the use of supporting software: *CLUMPP*, *distrupt* and *STRAT* are included, comprising the three most valuable tools used in the main applications of *STRUCTURE*. The user's guide offers a simplified view of how these tools can be applied to provide researchers with an informed choice of parameter settings and supporting software when analyzing their own genetic data.

43 An overview of the *STRUCTURE* program

45 *STRUCTURE* is a freely available program for population analysis developed by
46 Pritchard, Stephens and Donnelly, that was first described in 2000 (1). The program
47 analyses differences in the distribution of genetic variants amongst populations by
48 applying a Bayesian iterative algorithm that attempts to place a set of samples into
49 groups whose members share similar patterns of variation. *STRUCTURE* both identifies
50 populations from the data and assigns individuals to the population representing the best
51 fit for the patterns of variation found. Typically *STRUCTURE* is used as a first step to
52 examine the overall population structures that emerge from the sample set and this in
53 turn provides a preamble to further genetic analysis of the identified subpopulations or
54 can help to infer the origins of individuals with unknown population characteristics,
55 especially when population admixture has occurred. As *STRUCTURE* is based on the
56 core Bayesian principle of a comparison of likelihoods, there is the advantage that prior
57 information about the samples to be analysed can be supplied in a way that can shape
58 the analysis. For example, it is possible to input information about sampling location – a
59 characteristic that, when shared between individuals, might be associated with their
60 genetic proximity. The definition of populations in any species can be assessed from
61 geographical distribution, but is also often based on alternative criteria, including: the
62 phenotype, behaviour and ecology of the sampled individuals. Human populations can
63 also be defined by linguistic and cultural characteristics. Therefore it is important to
64 know whether a given assignment of individuals to populations based on non-genetic
65 criteria is consistent with differences in the genetic patterns detected between
66 populations (1-3).

67

68 The *STRUCTURE* algorithm uses a systematic Bayesian clustering approach where a
69 *Markov Chain Monte Carlo* (MCMC) estimation is applied. The MCMC process begins
70 by assigning individuals at random to a pre-determined number of groups. The variant
71 frequencies are then estimated in each group and individuals re-assigned based on those
72 frequency estimates. This is repeated a large number of times, typically comprising
73 100,000 iterations, in a process termed the burnin. The result of the MCMC burnin is a
74 progressive convergence towards reliable estimates of variant allele frequencies for each
75 population and probabilities of membership for each individual to a population.

76
77 Measurement of the assumed number of populations is performed separately from the
78 burnin process applying the MCMC estimation. *STRUCTURE* performs individual
79 analysis runs for each assumed population number from one up to a number reasonably
80 appropriate for the sampling regime. *STRUCTURE* applies a model to the data in which
81 there are K assumed populations or genetic groups, each of which is characterized by a
82 subset of allele frequencies identified in the data. Commonly K is not readily defined by
83 the user for the set of samples analysed, nevertheless it is still a parameter that must be
84 pre-selected. Therefore an appropriate first step in any population analysis using
85 *STRUCTURE* is to calculate the likelihood of the data for a range of K values by
86 creating posterior probabilities of K , known as X and written as X/K . It should be
87 emphasized that K is not an absolute value so user-defined values should be considered
88 carefully taking into account any recorded characteristics of the populations sampled.
89 Running a range of prescribed K settings to obtain their X values normally results in
90 several probabilities smaller than that obtained for the most appropriate K value but
91 beyond this point the probabilities tend to be very similar for higher K values. Therefore
92 plots of X values typically progress to a plateau for levels of K beyond the most

applicable number of detected populations such that the smallest of the stable K values represents the optimum value for the data. Kalinowski has noted that better clusters are created from the data if the most realistic K values are applied (4), so it is prudent to obtain the smallest value of K that maximizes the global likelihood of data – an approach which will capture the major underlying population structure in the data without overestimating it.

During each analysis run membership coefficients that sum to one are assigned to individuals for each group. The membership coefficient matrix, termed the *individual Q-matrix*, is generated and composed of rows corresponding to the number of individuals analysed and columns to the K number of clusters. The average of individual membership coefficients to each population forms the *population Q-matrix*. If admixture is not considered as a characteristic of the population samples analysed, the posterior probability for each individual of belonging to each of the K groups is calculated, and a sample can be considered a member of the group with the highest probability. If admixture is considered, the fraction of each individual's variation derived from each group is estimated and membership coefficients are made across multiple clusters.

Bayesian population analysis methods make a simple equation between the allele frequencies that define the population and the frequencies found in individuals identified as originating from that population. Therefore the ability of Bayesian methods to differentiate populations amongst a set of samples can be severely restricted when using limited sample sizes and small numbers of markers (5,6). The genetic markers that are applied to a *STRUCTURE* analysis are ideally: selectively neutral, showing reasonably low mutation rates and free of linkage disequilibrium (1,5). In short, they are assumed to be independent variables and Bayesian approaches assuming independence

(but without a guarantee this assumption is correct) are often termed naïve. However, enhancements since *STRUCTURE* version 2.3.1 allow the inclusion of markers that are weakly linked and therefore show some degree of non-independence (7).

Ancestry and allele frequency models

STRUCTURE implements different models of population structure to the data so selection of the most appropriate model depends on the user's data and study objectives. Therefore the guide that accompanies this review (Supplementary Material 1) centres on the effect changes to such models and prior population information can have on the results produced by *STRUCTURE*.

Two ancestry models applied by *STRUCTURE* are the *no admixture* and *admixture* models. If there is no prior knowledge about the origin of the populations under study or if there is reason to consider each population as completely discrete, the *no admixture* model is appropriate. However admixture between populations is a common characteristic so a large proportion of the sampled individuals can have recent ancestors from multiple populations. In these cases knowing the approximate median value of the ancestral population proportions for each individual and/or their populations of origin is a very useful part of the characterisation of the populations under study. In these cases the *admixture* model is more appropriate. Both models can be used with consideration for sampling location information by applying the prior model parameter: *LOCPRIOR* to the population model. This option can be used when there is additional sample-characteristic data available to the user, including: linguistic, geographical, cultural or phenotypic information. The *LOCPRIOR* parameter is particularly informative when there are weak population structure signals – a situation that can result from a reduced

number of markers, small sample sizes or due to close relationships between populations.

The third model parameter is *linkage*, based on the *admixture* model and this is designed to deal with admixture linkage disequilibrium (LD): the characteristic of extended LD found in admixed populations and often deliberately sought in association studies. This model was outlined by Falush et al. in 2003 (7) and provides more accurate estimates of statistical uncertainty when linked markers are used.

No admixture, *admixture* and *linkage* models can also be analysed as part of the *USEPOPINFO* model. This model uses the population labels to calculate the probability that each individual has to originate from the assumed population – individuals with low probabilities can be considered as hybrids or migrants. This parameter should be used with caution and applied only when the population labels are well defined beforehand and correspond almost exactly to the groups ultimately defined by the *STRUCTURE* results. The disadvantage of the *USEPOPINFO* model arises with the posterior handling of the results. The *individual Q-matrix* comprises probabilities (and not ancestry membership proportions) that are presented in a format that is incompatible with post-hoc data-processing software such as *CLUMPP* or *distruct*.

All the models considered until now can be used in conjunction with an alternative approach to *USEPOPINFO* – the *POPFLAG* model. *POPFLAG* considers the specified information about the population of origin of a portion of the individuals to help infer the ancestry of other samples with unknown origin. This option should also be used with caution because selected samples will be treated as the “reference” set (pre-assigned *POPFLAG*=1) so allele frequency estimates are based on a reduced sub-set of samples and will directly affect the grouping of the unknown individuals (pre-assigned *POPFLAG*=0). *POPFLAG* is an artificial model that assesses the individual probability

of being part of a particular population, but it can be useful if the objective is to efficiently group individuals/populations by comparison with a particularly well-defined and studied reference data set (1). One such reference set, widely applied to human population genetics studies, is the CEPH human genome diversity panel (HGDP-CEPH) (8) with the advantage that population structure has been identified in this sample set in a wide range of studies using different markers and a variety of data depths, but with consistent findings (9-11). When the *POPFLAG* model is used in conjunction with the *USEPOPINFO* model the *individual Q-matrix* is composed of two distinct parts: for *POPFLAG*=1 individuals the matrix presents probabilities, while for *POPFLAG*=0 individuals ancestry membership proportions are given according to the admixture model defined (*no admixture, admixture or linkage*).

Two allele frequency models are available. The *correlated allele frequencies* model assumes a level of non-independence, so is more conservative. The *independent allele frequencies* model requires knowledge about the correlation levels across populations – allele frequencies should be reasonably different in distinct populations. On the other hand, the *correlated allele frequencies* model provides greater power to detect distinct populations that are particularly closely related, although this model will give the same results as the *independent allele frequencies* model in the absence of high levels of correlation across populations. Therefore it is prudent to use the *correlated allele frequencies* model since this will guarantee that a previously undetected correlation is identified, but without affecting the results if no such correlation exists.

It is important to note that a level of finesse exists when implementing analysis models in *STRUCTURE*. All the models described above include specific statistical parameters that can be adjusted to more sensitive values, including r (informativeness of the

sampling location data), *alpha* (relative admixture levels between populations) and *lambda* (quantifies the independence between markers in terms of their allelic frequency distribution).

Assignment of individuals to a population and choice of markers

Assigning individuals to populations is often useful in population genetics studies (1) where making a population classification can provide an inference of individual ancestry that may not have been adequately defined beforehand (12). The typical approach has already been described for *STRUCTURE*: establishing pre-defined populations from reference samples and assigning individuals of unknown origin to these populations. Reference samples provide allele frequency estimates in each population that are then used to compute the likelihood of membership of samples of unknown origin to any population (1).

When using small numbers of markers, highly differentiated genetic variants are more informative per locus than randomly chosen markers. In these cases a measure of marker differentiation or divergence becomes an important factor in selecting markers to type. The informativeness metric I_n , proposed by Rosenberg *et al.* (13) is a useful measure of individual divergence per locus and per population comparison that can help guide marker choice ahead of commitments to the necessary genotyping.

The best markers for the inference of ancestry membership proportions are those that efficiently distinguish different populations, i.e. markers showing different alleles at very high frequency in distinct parental populations. Since fixed variation, private to one population, is very rare (14,15) marker selection must always be broadened to loci with maximized allele frequency differences between ancestral populations – these are

usually termed Ancestry Informative Markers or AIMs (11,16,17). Autosomal single nucleotide polymorphisms (SNPs) are increasingly favoured for human population analysis because, in addition to their widespread genomic distribution and ease of genotyping in very dense marker arrays, they are independent of admixture sex bias that routinely affects the distribution of variation of the Y chromosome or mitochondrial DNA. Segregating autosomal markers allow a more thorough measure of admixture in an individual contributed by all of their ancestors rather than just those of single uni-parental lineages (12,15,18,19).

Reference samples and variation databases

Amongst the objectives of human population genetics is the measurement of population-related parameters (e.g. effective size, degrees of relatedness, effects of local natural selection), the detection of admixture and the reconstruction of past demographic events. Therefore the proper definition of population structure is a key step in studying the populations of a region. In the case of admixed populations it is particularly important to define the original contributing populations by characterising reference populations and databases of human variation form the primary data sources for such studies.

A good starting point for collating human SNP variation data from the most extensive catalogues and for standard reference populations is *SPSmart* (<http://spsmart.cesga.es>, (20)). *SPSmart* has the advantage of being inclusive of all current SNP databases, specifically: 1000 Genomes, HapMap, Perlegen and Universities of the Stanford and Michigan CEPH-HGDP repositories. Additionally *SPSmart* allows the collection of genotype data from a large number of markers at a time and their direct (some data re-

arrangement is necessary to create the input file) transfer into population analysis programs of choice, including *STRUCTURE*.

The HGDP-CEPH is frequently used as a panel of population reference samples and CEPH panel samples from the same predefined population analysed with *STRUCTURE* nearly always share similar membership coefficients in inferred clusters (1,2,9-11).

Royal *et al.* noted that there are some limitations to the accuracy of ancestry inference within and among regions (12) that may be the result of the incomplete sampling by the CEPH-HGDP of total human genetic diversity (21). The first study using *STRUCTURE*

was performed in 2002 by Rosenberg *et al.* when the HGDP-CEPH worldwide population sample was analysed with 377 microsatellites to infer human population structure (10). This study concluded that the world's population could be grouped into six major discrete ancestral groups that match well with continental distributions.

Subsequent studies confirmed that when individuals are grouped on the basis of genetic similarity, group membership corresponds closely to predefined regional or population groups or to collections of geographically and linguistically similar populations (22,23).

In particular, the study of Li *et al.* in 2008, using *FRAPPE*, a very similar alternative population clustering method to *STRUCTURE*, divided the HGDP-CEPH into seven major population groups (23). Furthermore, such studies indicated it was also possible

to infer the ancestry of individuals from recently admixed populations in the context of the contributions of putative parental populations (17). Mixed ancestries inferred from genetic data can often be interpreted as arising from recent admixture among multiple founder populations. However, it can also be the result of a shared ancestry before the divergence of the two populations with a lack of subsequent gene flow between them (23).

268 Case-control association studies

269

270 Case-control association studies (CCAS) are a powerful strategy to identify loci that
271 contribute to complex diseases. The simplest approach involves the genotyping of a set
272 of markers in a sample of cases and unrelated controls, and then testing for allele
273 frequency differences at each marker – association of particular genomic regions
274 indicates that the loci they contain are possibly linked to disease susceptibility or to the
275 presence or absence of any particular phenotype (24). However the presence of
276 population structure between the case and the control groups can produce a high rate of
277 false positives due to allele frequency differences between subpopulations being
278 unrelated but mimicking allelic associations with the studied disease.

279 Amongst the best-known examples of the failure of association studies to record the
280 impact of population structure is the analysis of non-insulin-dependent diabetes in
281 Native American Pima and Papago populations. Results indicated that there was a
282 strong association between a high prevalence of diabetes and the absence of the
283 Gm3^{5,13,14} haplotype, but this actually resulted from the coincidental low frequency of
284 Gm3^{5,13,14} in the admixed Native American case subjects with diabetes contrasting with
285 a high frequency in European controls. The study authors suggested that Gm3^{5,13,14}
286 reduces the risk of diabetes, but the experimental design failed to allow for unrelated
287 frequency differences between the case and control groups (25). This is an example of
288 an incorrect attribution of a marker to a disease due to failure to detect and adjust for
289 stratification between case and control groups, leading to disease risk odds ratios for
290 unrelated loci that were likely much higher than the loci with real associations (26).

291 Two main approaches are favoured to overcome the effects of hidden or cryptic
292 structure when it exists between case and control groups: genomic control (GC) and

293 structured association (SA). Pritchard and Donnelly (24), reviewed genomic control
294 methods using the chi-square test to detect population stratification through the
295 estimation of the increase of the test statistic null distribution in comparison to the
296 distribution of values across unlinked markers typed in the same group. Using the
297 adjusted distribution it is possible to obtain corrected p -values at any given locus.
298 Structured association approaches use additional genotype information from unlinked
299 markers to estimate the number of subpopulations and each individual's assignment to
300 these subpopulations. This information can then be used to construct a test for
301 association (24). SA methods perform well but are computationally demanding and are
302 very reliant on estimating the correct number of subpopulations. In comparison, GC
303 methods while faster and more straightforward, can lack power in certain scenarios, for
304 example, when the markers used are not informative for population comparisons.
305 Finding procedures based on logistic regression that are flexible, computationally fast,
306 and easy to implement also provide good protection against the effects of cryptic
307 substructure, even though they do not explicitly model the population structure (27).
308 However if there is enough information for reliable estimation of the sub-population
309 data, the power and flexibility offered by SA approaches, facilitated by dedicated
310 software such as *STRAT*, makes them preferable to GC methods (24,28). *STRAT* is a
311 method that can be used in association mapping, enabling valid case-control studies
312 even in the presence of population structure. This method was first described by
313 Pritchard et al. in 2000 (29). The application of *STRAT* improves association studies as
314 it takes into account the confounding effects of population stratification through the use
315 of selected panels of ancestry informative markers. Several studies have been conducted
316 using *STRAT*-based stratification control, when there is doubt about the validity of the
317 associations found (i.e. that they are not spurious due to population stratification) or

when it is known that two or more populations are admixed (30,31). The accuracy of the inferences improves with sample size, number of loci, and the degree of divergence between populations (1). An instructive study by Campbell et al. (35) analyzed the efficacy of stratification control by constructing a case-control group based on adult height then measuring stratification with 111 random SNPs and 67 AIMs. Both SNP sets failed to detect stratification between case and controls but a single SNP in the gene LCT (rs4988235: LCT-13910C→T) with a frequency difference between north and south Europe co-incidental with average height difference across the same geographic distance was strongly associated. Re-matching case and control subjects into equivalent numbers of north and south Europeans in each entails the loss of this spurious association. In 2009, Price et al. developed an AIM panel that is able to distinguish between north-west and south-east European ancestry and when applied to the samples from Campbell's study stratification was detected (28). These studies demonstrate the importance of careful marker selection, especially when analysing closely related populations.

Other applications: forensic analysis

Forensic DNA analysis is a powerful tool in near-universal use as a core part of police investigations. DNA profiling has become particularly powerful when used in conjunction with a DNA database of offender profiles. But there are situations when no match is obtained. In those cases any information that can be obtained from the DNA becomes valuable (32). Information regarding the probable ancestry of an unknown offender may help to direct investigations towards a smaller group of suspects (32,33), likewise interest is growing in the prediction of externally visible characteristics such as

eye color (34). At the present time, microsatellite short tandem repeat (STR) typing can be considered a standard approach and still the method of choice in the forensic field, providing extremely high discrimination power for most problems of human identification (35,36). Several studies have concluded that STR profiles could also be used for ancestry inference (33,37-39), but will only have sufficient reliability when used with other AIMs such as SNPs (33). One applied example of the benefits of adding specialized marker sets to enhance ancestry analysis in forensic casework is the 11-M Madrid bomb attack investigations where results of a 34-SNP ancestry test were analysed with *STRUCTURE* to infer the probable ancestry of suspects indicated (40). One disadvantage of *STRUCTURE* highlighted by this study is the difficulty of analyzing single profiles of genotypes with this algorithm. An alternative online classifier termed *Snipper* (18), with a near identical Bayesian algorithm underlying the analyses it performs, gives likelihoods of membership to ancestry groups inferred from user input training sets as reference material.

Non-human applications: wildlife and agricultural population genetics

The analysis of populations using genetically defined clusters from *STRUCTURE* is of widespread interest in the study of many different species in the fields of molecular ecology and conservational genetics. One wildlife application recently described involved the forensic identification of source populations for illegally harvested animals (41). In agriculture *STRUCTURE* can be used to monitor artificial selection and help to identify animals that need to be kept in the breeding process in order to increase genetic variability in cattle, horses and buffaloes (42), plus the detection of recent origins or crossbreeding between breed groups (43). *STRUCTURE* has been employed to study

how human selection through breeding has shaped genetic variation of plant species including the tomato and bean (44). One useful application is to establish correlations between the population clusters detected in plant species and their ecotypes or patterns of geographical sowing (45). *STRUCTURE* has also been used to study the persistence of natural populations and assess the risk of escape of cultivated transgenic crops (46).

Alternative population analysis approaches

Several population analysis programs provide alternatives to *STRUCTURE* and are applicable to most of the analyses outlined above. Alternative approaches include:

- *ADMIXTURE*: a program for maximum likelihood estimation of individual ancestries from multilocus SNP genotypes. *ADMIXTURE* uses the same statistical model as *STRUCTURE* but calculates population parameters much more rapidly using a fast numerical optimization algorithm (47,48).

- *ADMIXMAP*: a program for modeling admixture, using marker genotypes and trait data on a sample of individuals from an admixed population, where the markers have been chosen to have strongly differentiated allele frequencies between two or more of the ancestral populations contributing to the admixture (49).

- *EIGENSOFT*: a program suite that has two main components, *EIGENSTRAT* uses principal component analysis to correct for population stratification in medical association studies (50) and *SMARTPCA* for the detection and analysis of population structure (51).

- *PLINK*: a program suite comprising a whole genome association analysis toolset designed to perform a range of basic, large-scale analyses in a computationally efficient manner (52).

393

394 Methods such as unrooted neighbour-joining trees, *Principal Component Analysis*
 395 (PCA) and *Multidimensional Scaling* (MDS) can also be informative to summarize the
 396 genetic similarities and differences between groups of populations (53). The review of
 397 Nassir et al. (54) suggests that PCA offers computational advantages if the markers are
 398 used for controlling population substructure in association studies. The suggested
 399 strategy of Nassir et al. is to use *STRUCTURE* for limiting analyses to particular subject
 400 groups, then apply PCA or MDS for association testing (54). PCA and MDS are also
 401 good alternatives to *STRUCTURE* for estimating the number of population clusters (51).
 402 While all the above programs were designed for a specific population analysis
 403 application, they lack the major advantage of *STRUCTURE* in offering the flexibility to
 404 adapt to different analysis demands.

405

406 The step-by-step guide to *STRUCTURE* analysis - an overview

407

408 In this article we present a review and practical guide for the use of the *STRUCTURE*
 409 population analysis program plus the associated software: *CLUMPP*, *distrupt* and
 410 *STRAT*. The user guide to *STRUCTURE* accompanying the article in Supplementary
 411 Material 1, comprises a step-by-step outline and covers the fundamentals of creating an
 412 input file and project, the available analysis models, the definition of parameter sets and
 413 how to run a simulation. The guide runs through an example where each of the analysis
 414 models and principal parameters of the four software tools are explored. The genotypic
 415 data used in this example is available in Supplementary Table 1. We present
 416 suggestions for the analysis and graphical display of the results and optimum estimation
 417 of the number of populations detected in a dataset.

418 Additionally, we describe how to handle the parameters included in *CLUMPP*, *distruct*
419 and *STRAT*. *CLUMPP* allows the alignment of different replicates of the *STRUCTURE*
420 analysis results of a given number of assumed populations, helping to deal with the
421 commonly encountered problem of multimodality. Clustering algorithms such as the
422 one implemented in *STRUCTURE* can include stochastic simulations during the
423 inferences. This creates a results space composed of different membership coefficients,
424 each one with an associated probability. It is possible that, when analysing the same
425 data set with identical conditions used, different final results are obtained. Differences
426 between replicated analysis runs can be of two types: *label change* or *genuine*
427 *multimodality*. Label change occurs when different replicates create the same
428 membership coefficient estimates but the labels of each group are distinct in each
429 permutation, that is, each cluster does not represent the same predefined population in
430 all runs. It is also possible that the replicates create distinct but likely results that are not
431 equivalent between permutations – genuine multimodality – that is, clusters that
432 represent a particular predefined population have different ancestry membership
433 proportions in each run. This can be the result of difficulties in the search for the
434 possible membership coefficients space or of true biological factors (55).

435 Independently of genuine differences between a series of replicated analyses, a method
436 is needed to deal with the replicate results obtained from multiple runs analysing a
437 single dataset. *CLUMPP* software provides three algorithms that identify the best
438 alignment to the replicate results of the cluster analysis. *CLUMPP* reviews the
439 membership coefficient matrices finding those replicates with the best correspondence.

440 Both *STRUCTURE* and *CLUMPP* give similar output files to the extent that *CLUMPP*
441 results can be directly used in standard *STRUCTURE* graphic enhancement software
442 such as *distruct* described below.

An informative way to visualize *STRUCTURE* results is to show each individual as a column segmented into K colours representing the estimated membership coefficients. *Distruct* software offers a wide range of options to create images based on the principal of segmented columns and provides visually appealing graphical summaries of the population structure detected in the data. *Distruct* is required by *CLUMPP* but usefully provides an alternative graphical display to the standard *STRUCTURE* bar plots, offering a wider variety of options to create images for much of the *STRUCTURE* output.

Finally, *STRAT* can be used in association studies, enabling the validation of case-control association statistics even in the presence of population structure that can compound the associations suggested by the data.

We focus this work on the current front-end version of *STRUCTURE* but it is worth noting that the recently released *StrAuto v0.3.1* Python-based software enables an automated approach, albeit from the command line of *Mac* or *Linux* based computers (56).

Concluding remarks

This article presents an up-to-date review of the ubiquitous *STRUCTURE* population analysis software widely applied to a range of population genetics problems. We give recommendations that can guide decisions when analyzing population structure for population genetics and association studies. The review and guide focuses on *STRUCTURE* and the supporting software of *CLUMPP*, *distruct* and *STRAT*. The use of a Bayesian method offers some advantages, especially assigning admixed individuals to population clusters since it is possible to use prior information to assist the calculation

of ancestry proportions for these individuals. Therefore information on data, the markers applied and the type of analysis desired is relevant before the selection of the analysis parameters for this Bayesian approach.

A simulated example file was thoroughly analyzed and our concluding remark is that there is not a standard analysis parameter in *STRUCTURE* – the data and the study objectives will influence the choice of the most appropriate parameter – and precautions should be adopted in order not to overestimate the population structure present on real and complex data.

Acknowledgements

LPH is supported by funding from Colciencias Colombia. YR is supported by funding from the Fundación *Gran Mariscal de Ayacucho* (FUNDAYACUCHO) E-228-585-2007-1. CS is supported by funding awarded by the Portuguese Foundation for Science and Technology (FCT) and co-financed by the European Social Fund (Human Potential Thematic Operational Programme (SFRH/BD/75627/2010). MVL is supported by funding from Xunta de Galicia INCITE 09208163PR. The authors wish to thank Antonio Salas of the Forensic Genetics Unit, University of Santiago de Compostela and Jeremy Austin of the Australian Centre for Ancient DNA, University of Adelaide, for helpful discussions and guidance in the preparation of the manuscript.

Competing Interests Statement

The authors declare no competing interests.

References

1. Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**:945-959.
2. Jobling, M.A., M. Hurles, and C. Tyler-Smith. 2004. *Human Evolutionary Genetics: Origins, Peoples & Disease*. Garland Science - Taylor & Francis Group, NY.
3. Waples, R.S. and O. Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol Ecol* **15**:1419-1439.
4. Kalinowski, S.T. 2011. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity* **106**:625-632.
5. Corander, J., P. Waldmann, and M.J. Sillanpaa. 2003. Bayesian Analysis of Genetic Differentiation Between Populations. *Genetics* **163**:367-374.
6. Corander, J. and P. Marttinen. 2006. Bayesian identification of admixture events using multilocus molecular markers. *Mol Ecol* **15**:2833-2843.
7. Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* **164**:1567-1587.
8. Cann, H.M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, et al. 2002. A Human Genome Diversity Cell Line Panel. *Science* **296**:261-262.

- 516 9. Abdulla, M.A., I. Ahmed, A. Assawamakin, J. Bhak, S.K. Brahmachari, G.C.
517 Calacal, A. Chaurasia, C.H. Chen, et al. 2009. Mapping Human Genetic
518 Diversity in Asia. *Science* *326*:1541-1545.
- 519 10. Rosenberg, N.A., J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A.
520 Zhivotovsky, and M.W. Feldman. 2002. Genetic structure of human
521 populations. *Science* *298*:2381-2385.
- 522 11. Enoch, M.A., P.H. Shen, K. Xu, C. Hodgkinson, and D. Goldman. 2006. Using
523 ancestry-informative markers to define populations and detect population
524 stratification. *J Psychopharmacol* *20*:19-26.
- 525 12. Royal, C.D., J. Novembre, S.M. Fullerton, D.B. Goldstein, J.C. Long, M.J.
526 Bamshad, and A.G. Clark. 2010. Inferring genetic ancestry: opportunities,
527 challenges, and implications. *Am J Hum Genet* *86*:661-673.
- 528 13. Rosenberg, N.A., L.M. Li, R. Ward, and J.K. Pritchard. 2003. Informativeness
529 of genetic markers for inference of ancestry. *Am J Hum Genet* *73*:1402-1422.
- 530 14. Pfaff, C.L., J. Barnholtz-Sloan, J.K. Wagner, and J.C. Long. 2004. Information
531 on ancestry from genetic markers. *Genet Epidemiol* *26*:305-315.
- 532 15. Lao, O., K. van Duijn, P. Kersbergen, P. de Knijff, and M. Kayser. 2006.
533 Proportioning whole-genome single-nucleotide-polymorphism diversity for the
534 identification of geographic population structure and genetic ancestry. *Am J*
535 *Hum Genet* *78*:680-690.
- 536 16. Salas, A., C. Phillips, and A. Carracedo. 2006. Ancestry vs physical traits: the
537 search for ancestry informative markers (AIMs). *Int J Legal Med* *120*:188-189.
- 538 17. Yang, N., H.Z. Li, L.A. Criswell, P.K. Gregersen, M.E. Alarcon-Riquelme, R.
539 Kittles, R. Shigeta, G. Silva, et al. 2005. Examination of ancestry and ethnic
540 affiliation using highly informative diallelic DNA markers: application to

- 541 diverse and admixed populations and implications for clinical epidemiology and
542 forensic medicine. *Hum Genet* **118**:382-392.
- 543 18. Phillips, C., A. Salas, J.J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-
544 Dios, M. Calaza, M.C. de Cal, et al. 2007. Inferring ancestral origin using a
545 single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int*
546 *Genet* **1**:273-280.
- 547 19. Halder, I., M. Shriver, M. Thomas, J.R. Fernandez, and T. Frudakis. 2008. A
548 panel of ancestry informative markers for estimating individual biogeographical
549 ancestry and admixture from four continents: Utility and applications. *Hum*
550 *Mutat* **29**:648-658.
- 551 20. Amigo, J., A. Salas, C. Phillips, and A. Carracedo. 2008. SPSmart: adapting
552 population based SNP genotype databases for fast and comprehensive web
553 access. *Bmc Bioinformatics* **9**:-.
- 554 21. ASHG. 2008. The American Society Of Human Genetics: Ancestry Testing
555 Statement.
- 556 22. Allocco, D.J., Q. Song, G.H. Gibbons, M.F. Ramoni, and I.S. Kohane. 2007.
557 Geography and genography: prediction of continental origin using randomly
558 selected single nucleotide polymorphisms. *BMC Genomics* **8**:68.
- 559 23. Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S.
560 Ramachandran, H.M. Cann, G.S. Barsh, et al. 2008. Worldwide human
561 relationships inferred from genome-wide patterns of variation. *Science*
562 **319**:1100-1104.
- 563 24. Pritchard, J.K. and P. Donnelly. 2001. Case-control studies of association in
564 structured or admixed populations. *Theor Popul Biol* **60**:227-237.

- 565 25. Knowler, W.C., R.C. Williams, D.J. Pettitt, and A.G. Steinberg. 1988.
566 Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians
567 with genetic admixture. *Am J Hum Genet* 43:520-526.
- 568 26. Cardon, L.R. and L.J. Palmer. 2003. Population stratification and spurious allelic
569 association. *Lancet* 361:598-604.
- 570 27. Setakis, E., Stirnadel, H., Balding, D.J.,. 2005. Logistic regression protects against
571 population structure in genetic association studies. *Genome Res*:292-296.
- 572 28. Price, A.L., J. Butler, N. Patterson, C. Capelli, V.L. Pascali, F. Scarnicci, A.
573 Ruiz-Linares, L. Groop, et al. 2008. Discerning the ancestry of European
574 Americans in genetic association studies. *PLoS Genet* 4:e236.
- 575 29. Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. 2000. Association
576 Mapping in Structured Populations. *Am J Hum Genet* 67:11.
- 577 30. Han, S., Guthridge, J.M., Harley, I.T.W., Sestak, A.L., Kim-Howard, X.,
578 Kaufman, K.M., Namjou, B., Deshmukh, H., Bruner, G., Espinoza, L.R.,
579 Gilkeson, G.S., Harley, J.B., James, J.A., and Nath, S.K. 2008. Osteopontin
580 and Systemic Lupus Erythematosus Association: A Probable Gene-Gender
581 Interaction. *PLoS One* 3.
- 582 31. Tian, C., Plenge, R.M., Ransom, M., Lee, A., Villoslada, P., Selmi, C.,
583 Klareskog, L., Pulver, A.E., Qi, L., Gregersen, P.K., and Seldin M.F. 2008.
584 Analysis and Application of European Genetic Substructure Using 300 K SNP
585 Information. *Plos Genet* 4:4.
- 586 32. Lowe, A.L., A. Urquhart, L.A. Foreman, and I.W. Evett. 2001. Inferring ethnic
587 origin by means of an STR profile. *Forensic Sci Int* 119:17-22.
- 588 33. Phillips, C., L. Fernandez-Formoso, M. Garcia-Magarinos, L. Porras, T.
589 Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, et al. 2011. Analysis

- 590 of global variability in 15 established and 5 new European Standard Set (ESS)
 591 STRs using the CEPH human genome diversity panel. *Forensic Sci Int Genet*
 592 *5*:155-169.
- 593 34. Ruiz, Y., C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. Casares de Cal, R.
 594 Cruz, O. Maroñas, J. Sochtig, et al. 2012. Further development of forensic eye
 595 color predictive tests. *Forensic Sci Int: Genet* doi:10.1016/j.fsigen.2012.05.009.
- 596 35. Chakraborty, R., D.N. Stivers, B. Su, Z. Yixi, and B. Bruce. 1999. The utility of
 597 short tandem repeat loci beyond human identification: Implications for
 598 development of new DNA typing systems. *Electrophoresis* *20*:1682-1696.
- 599 36. Butler, J.M. 2005. *Forensic DNA typing*. Burlington, London Elsevier Academic
 600 Press.
- 601 37. Graydon, M., F. Cholette, and L.K. Ng. 2009. Inferring ethnicity using 15
 602 autosomal STR loci--comparisons among populations of similar and distinctly
 603 different physical traits. *Forensic Sci Int Genet* *3*:251-254.
- 604 38. Londin, E.R., M.A. Keller, C. Maista, G. Smith, L.A. Mamounas, R. Zhang,
 605 S.J. Madore, K. Gwinn, and R.A. Corriveau. 2010. CoAIMs: a cost-effective
 606 panel of ancestry informative markers for determining continental origins. *PLoS*
 607 *One* *5*:e13443.
- 608 39. Bowcock, A.M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L.
 609 Cavalli-Sforza. 1994. High resolution of human evolutionary trees with
 610 polymorphic microsatellites. *Nature* *368*:455-457.
- 611 40. Phillips, C., L. Prieto, M. Fondevila, A. Salas, A. Gomez-Tato, J. Alvarez-Dios,
 612 A. Alonso, A. Blanco-Verea, et al. 2009. Ancestry analysis in the 11-M Madrid
 613 bomb attack investigation. *PLoS One* *4*:e6583.

- 614 41. Ball, M.C., L.A. Finnegan, T. Nette, H.G. Broders, and P.J. Wilson. 2011.
615 Wildlife forensics: "Supervised" assignment testing can complicate the
616 association of suspect cases to source populations. *Forensic Sci Int Genet* 5:50-
617 56.
- 618 42. Gargani, M., L. Pariset, M.I. Soysal, E. Azkan, and A. Valentini. 2010. Genetic
619 variation and relationships among Turkish water buffalo populations. *Anim*
620 *Genet* 41:93-96.
- 621 43. Martn-Burriel, I., C. Rodellar, J. Cann, O. Corts, S. Dunner, V. Landi, A.
622 Martnez-Martnez, L.T. Gama, et al. 2011. Genetic diversity, structure, and
623 breed relationships in Iberian cattle. *J Anim Sci* 89:893-906.
- 624 44. Kwak, M. and P. Gepts. 2009. Structure of genetic diversity in the two major gene
625 pools of common bean (<i>Phaseolus vulgaris</i> L., Fabaceae).
626 *Theor Appl Genet* 118:979-992.
- 627 45. Santalla, M., A. De Ron, and M. De La Fuente. 2010. Integration of genome and
628 phenotypic scanning gives evidence of genetic structure in Mesoamerican
629 common bean (*Phaseolus vulgaris* L.) landraces from the southwest of Europe.
630 *Theor Appl Genet* 120:1635-1651.
- 631 46. Pascher, K., S. Macalka, D. Rau, G. Gollmann, H. Reiner, J. Glossl, and G.
632 Grabherr. 2010. Molecular differentiation of commercial varieties and feral
633 populations of oilseed rape (*Brassica napus* L.). *BMC Evol Biol* 10:63.
- 634 47. Alexander, D.H., J. Novembre, and K. Lange. 2009. Fast model-based estimation
635 of ancestry in unrelated individuals. *Genome Res* 19:1655-1664.
- 636 48. Zhou, H., D.H. Alexander, and K. Lange. 2011. A quasi-Newton acceleration for
637 high-dimensional optimization algorithms. *Stat Comput* 21:261-273.

- 638 49. McKeigue PM, C.J., Parra EJ, Shriver MD:. 2000. Estimation of admixture and
639 detection of linkage in admixed populations by a Bayesian approach: application
640 to African-American populations. *Ann Hum Genet* **64**:171-186.
- 641 50. Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D.
642 Reich. 2006. Principal components analysis corrects for stratification in
643 genome-wide association studies. *Nat Genet* **38**:904-909.
- 644 51. Patterson, N., A.L. Price, and D. Reich. 2006. Population structure and
645 eigenanalysis. *PLoS Genet* **2**:e190.
- 646 52. Purcell S, N.B., Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J,
647 Sklar P, de Bakker PI, Daly MJ, Sham PC: . 2007. PLINK: a tool set for
648 whole-genome association and population-based linkage analyses. *Am J Hum*
649 *Genet* **81**:559-575.
- 650 53. Gao, X. and J. Starmer. 2007. Human population structure detection via
651 multilocus genotype clustering. *BMC Genet* **8**:34.
- 652 54. Nassir, R., R. Kosoy, C. Tian, P. White, L. Butler, G. Silva, R. Kittles, M.
653 Alarcon-Riquelme, et al. 2009. An ancestry informative marker set for
654 determining continental origin: validation and extension using human genome
655 diversity panels. *BMC Genet* **10**:39.
- 656 55. Jakobsson, M. and N.A. Rosenberg. 2007. CLUMPP: a cluster matching and
657 permutation program for dealing with label switching and multimodality in
658 analysis of population structure. *Bioinformatics* **23**:1801-1806.
- 659 56. Chhatre, V.E. January 3, 2012. StrAuto: A python Program - Automation of
660 Structure Analysis for Mac & Linux version 0.3.1.
- 661

Bloque 2.

***V.8. Assessing the forensic potential of an
eye colour predictive test in challenging
DNA***

Y. Ruiz, A. Freire-Aradas, M. Fondevila, C. Phillips, M.V. Lareu.

(Article in preparation, this work belongs to a further forensic validation project)

Assessing the forensic potential of an eye colour predictive test in challenging DNA

Y. Ruiz¹, A. Freire-Aradas¹, M. Fondevila¹, C. Phillips¹, M.V. Lareu¹.

Forensic Genetics Unit, Institute of Forensic Sciences “Luis Concheiro”, University of Santiago de Compostela, Spain.

Abstract

Prediction of external visible characteristics (EVCs) is becoming an important tool in Forensic Genetics. Several Single Nucleotide Polymorphisms (SNPs) have been associated in a number of pigmentation genes with various human hair, skin and eye colour phenotypes. Recently, it has been proposed the use of a predictive test for eye colour inference, trying to found an equilibrium between the predictive power for a range of eye colour and a moderated number of markers employed. In the present work, we evaluate the forensic potential of an eye colour assay proposed previously (Ruiz et al. Submitted, 2012), in terms of stability, reproducibility and the ability in detection for challenging biological material, obtained from cadaveric remains presenting high levels of inhibitors and a barely detection with individual identification markers employed in forensic casework.

1. Introduction

In the forensic field, choosing a genotyping technology that can be successfully applied to challenging DNA is an essential step. Several studies have evaluated the forensic potential of predictive assays based in SNPs for individual identification and ancestry inference [1-5]. Regarding degraded DNA, SNPs become a tempting alternative to currently used Short Tandem Repeats (STRs), mostly due to the feasible amplification of short-amplicons during the PCR. An adequate technique to detect low template DNA, as frequently found in forensic casework, should be sensitive and allow for multiplexing [5].

Inference of external visible characteristics (EVCs) has been a recent issue of interest in forensic human identification. Analysis of SNPs for prediction of EVCs is fundamental, not only because of the forensic potential, but due to the unique

associations found in SNPs with physical traits. Nowadays, the forensic community is attempting the development of new forensic tests for inference of EVCs, since prediction of these characteristics represent an important complement either in criminal investigations or human identification in mass disasters. In this sense, there have been published studies proposing the use of a predictive test for blue and brown eye colour determination [6-7], however presenting a low sensitivity for predicting non-blue and non-brown phenotypes [8-10]. Recently, it has been proposed the expansion in number of makers employed in eye colour prediction, in order to improve the classification of complex colours, since it has been explored the genetic regulation involved in this trait, trying to found an equilibrium between the predictive power for a range of eye colour and the ability in detection for challenging biological material. In the present work, we evaluate the forensic potential of an eye colour assay proposed in a previous study [11] applied in a challenging sample, obtained from a cadaveric remain, that showed high levels of inhibition and a barely detection with STRs.

2. Material and methods

Skeletal remains (from an humerus and a teeth) with twenty years of antiquity were analyzed. DNA extraction was achieved by a previously adapted phenol/chlorophorm method (Fondevila et al, 2008). DNA concentrations were determined using the Quantifiler™ Human DNA Quantification Kit with the AB 7300 real-time PCR systems (Applied Biosystems: AB). STR typing was performed by AmpFISTR Identifier®(AB), PowerPlex16 (Promega). MiniSTRs were tested by AmpFISTR MiniFiler(AB). The pigmentation SNP typing comprise the 23 eye colour associated markers contained in SHEP 1 (12) and SHEP 2 (11) assays [11]. Both sets use an initial PCR followed by two primer extension assays. For typing assays were used AB SNaPShot primer extension assay following protocols adapted from Ruiz et al. [11] SNP assays used undiluted DNA samples throughout. A serial dilution assay was performed with the SNP system using a positive control in order to test the stability of the test.

3. Results

The concentration value of the DNA extract from the humerus obtained was 0.00183 ng/μL, with a IPC of 30.45 which indicates the presence of inhibitors in the extract.

Therefore, the humerus represents an interesting sample for test the SNP-pigmentation assay, since it corresponds to a challenging material in comparison with the teeth, which showed no significant inhibition. The performance of the three STRs systems in the humerus extract showed negative profiles for AmpFISTR MiniFiler and PowerPlex16, and a barely detection obtained from AmpFISTR Identifier®, while for the two SNPs assay (SHEP1 and SHEP2) the success rate was 100 and 90% respectively. Table 1 outlines the genotyping success for DNA extract from humerus, employing SNP and STRs systems. The respective electropherograms from these assays are showed in Figure 1.

PCRs for SNP typing were performed in triplicate in order to confirm results. All markers with exception of two of them, were reproducible. These two markers (rs16891982 and rs4778138) have presented differences in terms of allele drop-out or alle drop-in, a common event when dealing with compromised samples, so it was not possible to obtain a consensus result for this two markers.

During the serial dilutions realized, both assays (SHEP1 and SHEP2) perform in a concentrations that ranges from 32,8 to 0,065 ng/ μ L (supplementary material 1). These concentrations were employed based in the experience treating with real forensic samples that usually stay into this margin. Moreover as expected, a diminution of the intensity (RFU) of the signal directly proportional to the dilution of the sample was observed, although it was still possible to obtain complete profiles till 0,656 ng/ μ L with a signals of 1000 RFU approximately.

4. Conclusions

A genetic test with forensic proposes, besides their accuracy in function for an individual identification, prediction of ancestry, or external visible characteristics, should owns the ability to be sensitive in presence of challenging DNA. The SNP-eye colour test SHEP1 and SHEP2 whose predictive power has been evaluated in a previous work, has demonstrated in the present study its forensic capability performance regarding a sample presenting either low template DNA or high level of inhibition. Additionally, the evaluation of the reproducibility and stability in this test suggest its ability to be used in future studies of forensic validation, that includes a further rigorous evaluation analysis.

References

- [1] Borsting C, Rockenbauer E, Morling N: Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. *Forensic Sci Int Genet* 2009, 4:34-42.
- [2] Fondevila M, Phillips C, Naveran N, Fernandez L, Cerezo M, Salas A, Carracedo A, Lareu MV: Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur. *Forensic Sci Int Genet* 2008, 2:212-218.
- [3] Phillips C, Prieto L, Fondevila M, Salas A, Gómez-Tato A, Álvarez-Dios J, Alonso A, Blanco-Verea A, Brión M, Montesino M: Ancestry analysis in the 11-M Madrid bomb attack investigation. *PloS one* 2009, 4:e6583.
- [4] Sanchez JJ, Phillips C, Borsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N: A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 2006, 27:1713-1724.
- [5] Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A: Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet* 2007, 1:273-280.
- [6] Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M: IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet* 2011, 5:170-180.
- [7] Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN: Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. *Forensic Sci Int Genet* 2011, 5:464-471.
- [8] Purps J, Geppert M, Nagy M, Roewer L: Evaluation of the IrisPlex eye colour prediction tool in a German population. *FSI: Genetic Supplement Series* 3 2011, e202-e203.
- [9] Prestes PR, Mitchell RJ, Daniel R, Ballantyne KN, Oorschot RAHv: Evaluation of the Irisplex system in admixed individuals. *FSI: Genetic supplement series* 2011, 3:e283-e284.
- [10] Bubul O, Filoglu G, Altuncul H, Freire-Aradas A, Ruiz Y, Fondevila M, Phillips C, Á C, Kriegel AK, Schneider PM: A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. *FSI: Genetic Supplement Series* 2011, e500-e501.
- [11] Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maroñas O, Söchtig J, Fondevila M, Rodriguez-Cid MJ, Carracedo Á, Lareu MV: Further development of forensic eye color predictive tests. *Forensic Science International: Genetics* 2012, doi:10.1016/j.fsigen.2012.05.009.

Table 1. Genotyping success of challenging DNA extracted from humerus, employing the SNPs eye colour test and the STRs identification assay.

Genotyping Assay	Typing success (%)	Marker type
SHEP 1	100	SNPs
SHEP 2	90,9	
<i>AmpFISTR Identifier®</i>	18	STRs
<i>PowerPlex®16 System</i>	0	
<i>AmpFISTR MiniFiler®</i>	0	

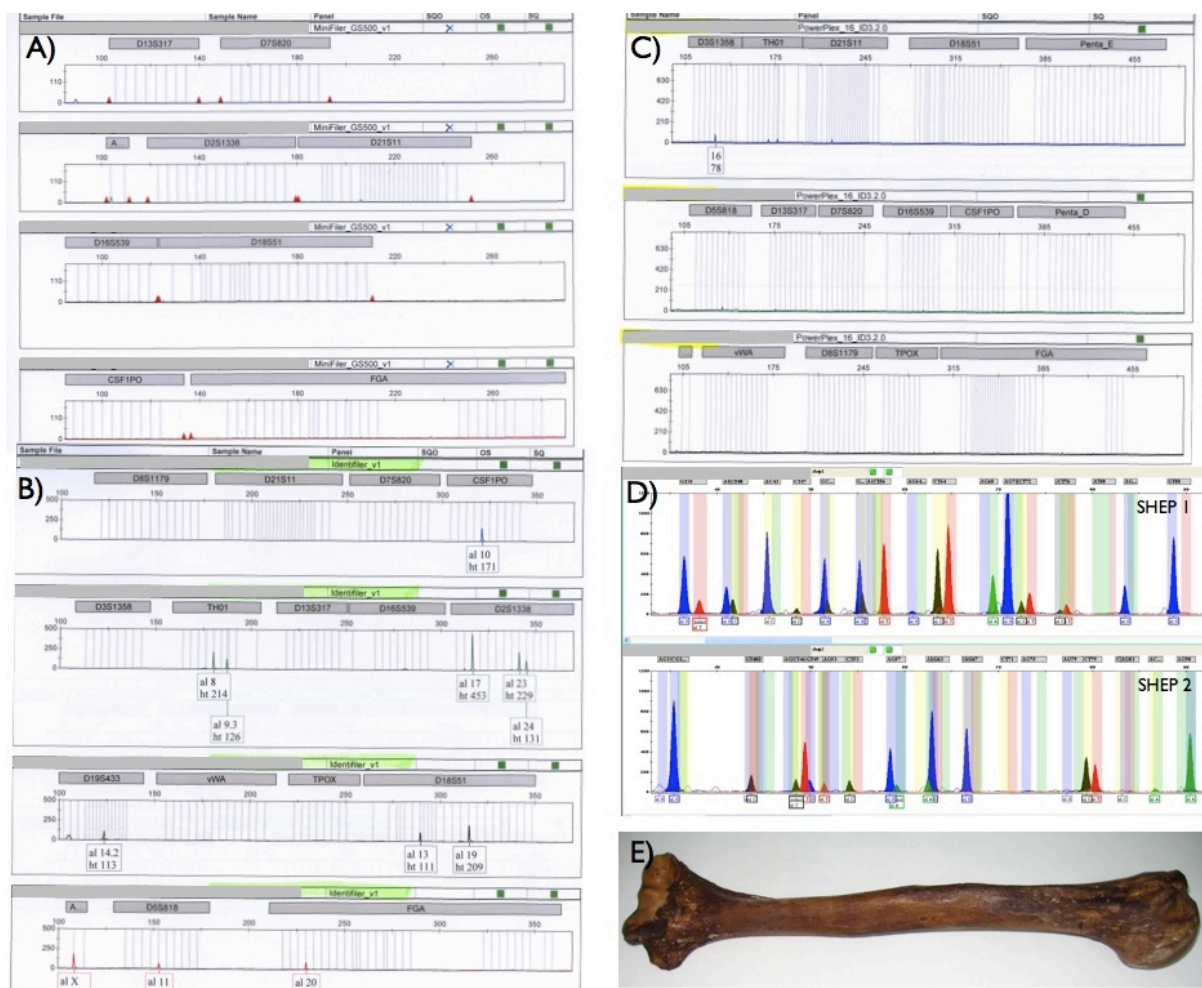
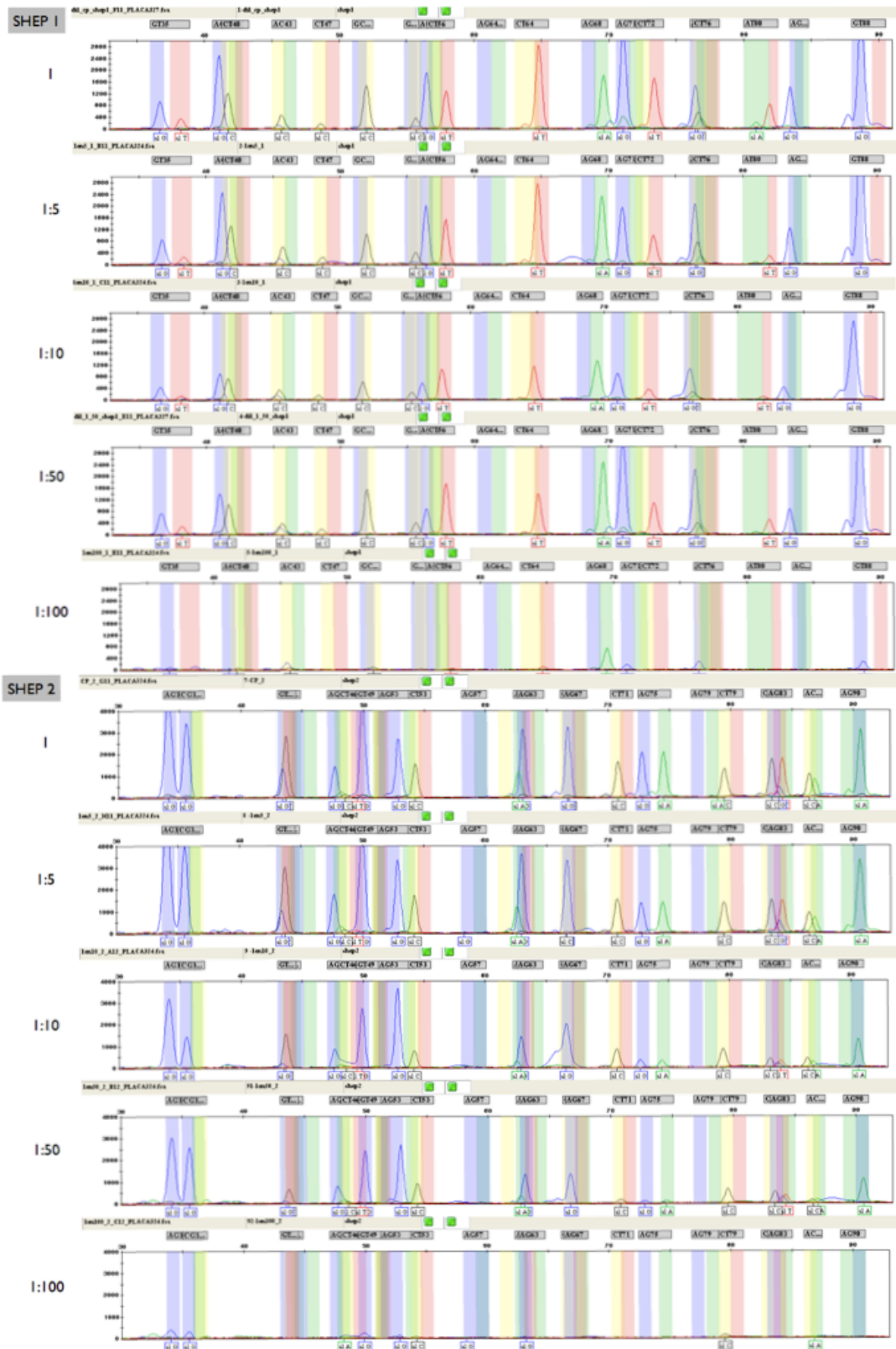


Figure 1. Electropherograms from humerus extract showing STR genotyping assays: *AmpFISTR MiniFiler®* (A), *AmpFISTR Identifier®* (B), *PowerPlex®16 System* (C), and SNP eye colour genotyping assays SHEP1/SHEP2 (D). Appearance of the femur employed (E).



Supplementary material 1. Serial dilution assays employing a positive control with an initial concentration of 32,8 ng/μL.

Bloque 2- Discusión

El empleo de marcadores genéticos en la predicción de una característica visible externa como la pigmentación humana, puede ser de gran interés en los siguientes escenarios demográficos:

- a. Las poblaciones con una amplia distribución y variabilidad del rasgo, que posean una ascendencia generalmente determinada por un grupo ancestral específico. Tal es el caso de los grupos de poblaciones europeas analizadas en el presente estudio, las cuales presentaron una amplia variabilidad en la pigmentación, como la observada en color de ojos.
- b. Las poblaciones que presentan una mezcla notoria en su composición ancestral, en las cuales la variabilidad del rasgo podría no estar definida. Por ejemplo, algunos grupos afroamericanos en USA, mestizos en Sudamérica, entre otros, que presentan una ascendencia compuesta por dos o más poblaciones ancestrales. Uno de los grupos que probablemente ilustre mejor este caso es Brasil. Como objetivo futuro, se pretende realizar un análisis conjunto de inferencia de grupos ancestrales y pigmentación en algunas poblaciones de esta región, cuyos patrones de mezcla resultan de gran interés para la evaluación de EVCs *vs* AIMs.

En ambos escenarios, el empleo de EVCs puede llegar a ofrecer información adicional en la investigación forense, criminalística y antropológica. Sin embargo, en aquellas poblaciones que manifiesten una variabilidad menor de pigmentación, además de pertenecer a un grupo ancestral definido como pueden ser por ejemplo, algunas poblaciones nativas de África, América, Asia y Oceanía, la predicción de EVCs podría no ser tan informativa como la inferencia del BGA, dependiendo de lo que sea interrogado. Por lo tanto, en presencia de un caso forense en el que se indague sobre posibles características adicionales de un individuo (además de los análisis de identificación de rutina correspondientes), es recomendable realizar un estudio preliminar de AIMs.

A pesar de que el contexto de la investigación forense pudiera corresponder a una población geográfica específica, resulta evidente que esto no determina la ascendencia que pudiera poseer el individuo en cuestión, más aún en la actualidad, en donde el flujo de migraciones entre individuos de diversas poblaciones ocurre constantemente, y donde además existe una mezcla manifestada entre estas poblaciones. Posterior al análisis de AIMs, dependiendo de los resultados obtenidos, y nuevamente, de lo que sea interrogado en cada caso, sería entonces recomendable aplicar el análisis de EVCs.

Por otra parte, es de interés recordar la distinción que existe entre la asociación de marcadores genéticos a un tipo de EVC como la pigmentación, y el poder de estos marcadores en la inferencia de grupos ancestrales. Un AIM, no necesariamente se relaciona con la expresión de un rasgo físico. De hecho, cualquier marcador considerado “neutral” puede ser empleado en este contexto para inferir grupos ancestrales. Sin embargo, no se descarta que el estudio de SNPs asociados a rasgos físicos pudiera ser una manera útil de detectar AIMs. El poder de predicción basado en marcadores genéticos, dependerá de cuan informativos sean estos, y de la historia de la distribución de sus alelos entre las diversas poblaciones. En el caso del estudio presentando sobre el análisis preliminar de grupos ancestrales eurasiáticos (V.6), basado en paneles de AIMs como el 34-plex, incluye algunos marcadores de relevancia funcional en la determinación de pigmentos como por ejemplo rs12913832 y rs16891982 los cuales han resultado ser de utilidad en la diferenciación de las poblaciones geográficas de estudio, posiblemente debido a eventos de selección positiva que han podido ocasionar una variación entre las frecuencias alélicas de estos SNPs en las poblaciones. Sin embargo, los análisis correspondientes a la inferencia del color de ojos fueron realizados de manera independiente a los de predicción de grupos ancestrales, a pesar de compartir algunos marcadores informativos.

Una de las principales limitaciones que enfrenta la investigación de EVCs, es la asignación de clases “fenotípicas” a características que en realidad son de tipo continuas. En el caso del estudio presentado sobre el desarrollo de un *test* para la inferencia del color de ojos, este rasgo a diferencia del cabello y piel, manifiesta una serie de patrones estructurales incluso de diversa coloración, lo cual aumenta la

complejidad al intentar categorizar (Fig. D2). Pese a las propuestas que se han realizado en busca de una clasificación de color de ojos adecuada, tanto en la bibliografía como en el presente trabajo de investigación, sólo la tonalidad (claro-oscuro) figura como una de las mejores aproximaciones a realizar en la predicción a partir de los marcadores genéticos que se conocen asociados, además de la predicción del color de ojos azules. Dada la complejidad, más allá de la variabilidad en la coloración, una inferencia óptima del color de ojos debería no sólo abarcar aquellos marcadores genéticos relacionados a la pigmentación, sino también a todos los polimorfismos que presenten una asociación significativa con la expresión de otras proteínas, las cuales aunque de manera indirecta también afectan la apariencia visual del ojo, mediante la manifestación de patrones estructurales en el iris como puede ser el grosor de éste, la presencia de anillos peripupilares, criptas de Fuch, presencia de colágeno, surcos de contracción, etc (Fig. D2). Aunque en la actualidad sólo se conocen los genes asociados a la expresión de algunas de estas proteínas en el iris, es muy probable que el siguiente paso en la investigación sobre la apariencia visual del ojo, sea el detectar e incorporar en un *test* de predicción, aquellos SNPs responsables de la expresión de estos patrones complejos, los cuales permitirán entonces realizar una inferencia más eficaz sobre este rasgo.

Otro de los factores a considerar en la inferencia de EVCs a partir de marcadores genéticos, es el modelo de predicción matemático escogido. La probabilidad condicionada de Bayes implementada en el clasificador *Snipper*, representa una alternativa a otros métodos, ya que además de las propiedades de su algoritmo (basado en ratios de verosimilitud), la aplicación web diseñada permite conjugar la información de diferentes marcadores mediante el clasificador por frecuencias, lo cual resulta de interés en el caso de la predicción de color de ojos, dada la presencia de diplotipos definidos en la región OCA2-HERC2, que han demostrado contribuir en la diferenciación de grupos complejos, además de los SNPs individuales previamente descritos.

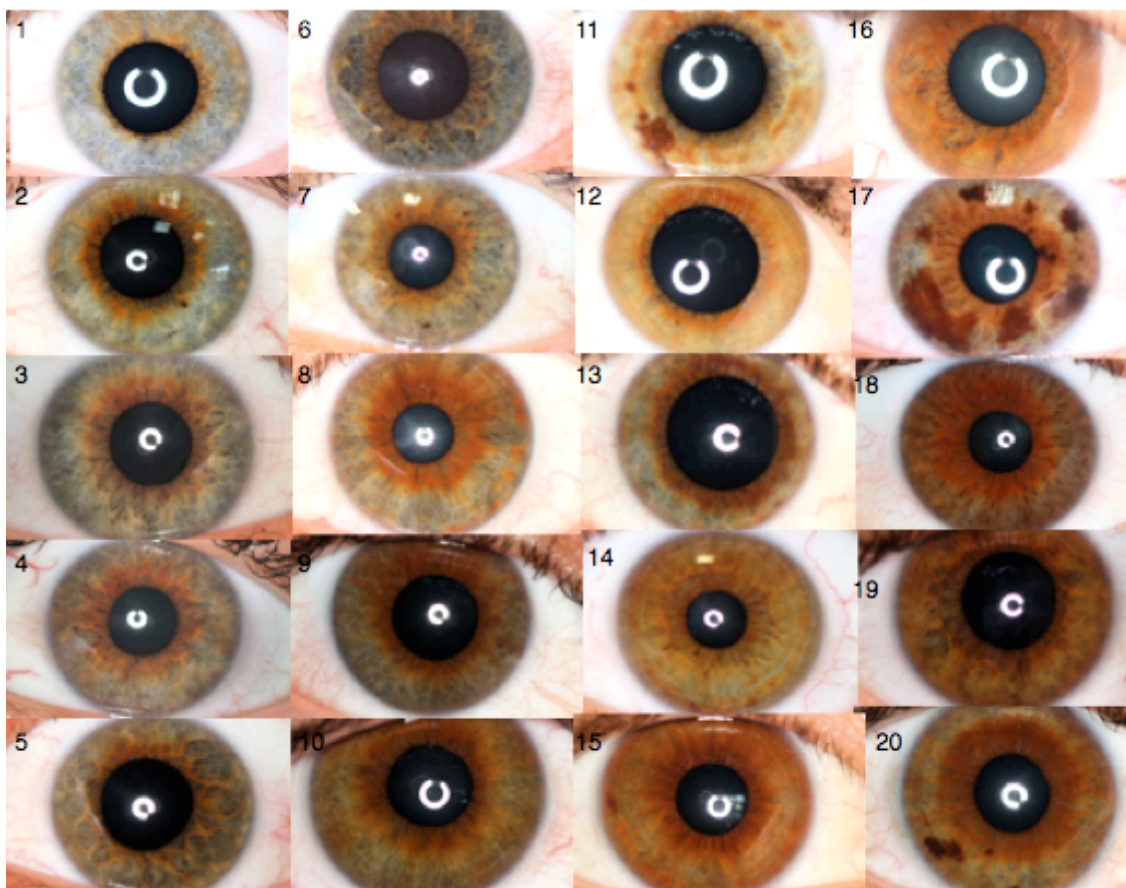


Fig. D2. Ejemplos de algunos patrones complejos encontrados en el iris: anillos peripupilares (2,8,13,18,20), criptas de Fuch (5,6,16,19) y manchas de melanina (11,17,20), surcos de contracción (10).

Sin embargo, una limitación tanto de este como de otros clasificadores propuestos en la bibliografía, es que no incorporan al modelo el efecto de las interacciones que se han descrito entre diversos marcadores genéticos, mediante la aplicación de algoritmos que computen dichos efectos. Esto, podría representar una mejora en la clasificación del color de ojos, lo cual también supone un reto futuro en la investigación tanto de éste, como de otros rasgos físicos complejos.

Por otra parte, la aplicación del programa *STRUCTURE* al estudio de color de ojos, representó una herramienta de gran valor en la compleja definición de las poblaciones de referencia que fueron empleadas. Hecho que demuestra el potencial y versatilidad tanto en ajustes, como en modelos que presenta este programa para el estudio de poblaciones, pese a la complejidad que pueda estar asociada a ellas. De igual forma, esta herramienta es de gran utilidad en la clasificación de individuos y

en la detección de posible mezcla genética que estos puedan manifestar, tal y como fue observado en los individuos de color de ojos verde-avellana.

Además de las recomendaciones expuestas anteriormente sobre los factores que podrían contribuir en un futuro a la mejora en la predicción de fenotipos complejos, actualmente se está realizando un proyecto piloto sobre los patrones de herencia en este rasgo, evaluando los principales marcadores asociados, especialmente los contenidos en la región OCA2-HERC2. Uno de los objetivos principales de este trabajo, será confirmar la posible independencia alélica observada en el estudio previo (V.5) entre los marcadores contenidos dentro de esta región, ya que en algunos de estos SNPs se ha observado un aparente patrón de segregación en bloque, en poblaciones de color de ojos no complejos principalmente, sin embargo, en presencia de fenotipos complejos (no-azul/no-marrón), tienden a manifestar un patrón de segregación independiente, es decir donde ocurre recombinación, aunque en una baja tasa. Por lo que a través de un estudio de herencia genética en familias, sería posible entonces confirmar los patrones de segregación observados, así como determinar los marcadores dentro de esta región que podrían intervenir en la determinación de fenotipos complejos.

En cuanto a su aplicación forense, el diseño del sistema *multiplex* para la inferencia de pigmentación humana ha sido realizado seleccionando marcadores genéticos informativos sobre el rasgo, que además resultaran adecuados para el análisis de muestras difíciles (*challenging DNA*), las cuales usualmente poseen un alto grado de inhibición y degradación, tal y como se mostró en el análisis de restos cadavéricos presentado. En este sentido, los SNPs son marcadores genéticos idóneos para este tipo de análisis, tomando en consideración que además son los únicos marcadores donde se han encontrado asociaciones a EVCs, a diferencia de los Indels y STRs, para los cuales no se ha descrito ninguna asociación que aporte la información necesaria sobre la inferencia de rasgos físicos hasta la fecha.

El *multiplex* SHEP además de haber sido probado en muestras obtenidas a partir de restos cadavéricos como húmero y diente, también se ha empleado con éxito en otras muestras biológicas a partir de semen, saliva, y manchas de sangre. En

un caso en particular, la operación Minstead (previamente mencionada), era de interés conocer además del grupo ancestral, la información sobre la pigmentación del agresor, por lo que se empleó para ello los multiplexes SHEP 1 y SHEP 2, obteniendo un perfil cuya verosimilitud era mayor para la predicción fenotípica de pigmentación de ojos marrón, así como de coloración de piel oscura, características que posteriormente fueron confirmadas con éxito tras la aprensión del agresor.

Por otra parte, la metodología de genotipado también representa un factor clave en desarrollo de un *test* con aplicación forense. El genotipado por *SNapShot*, aunque sigue siendo una de las tecnologías más empleadas en la detección de SNPs en este contexto, presenta una serie de limitaciones como la aparición de artefactos ocasionados por el empleo de pasos múltiples de *post-pcr*, lo que resulta aún más crítico en presencia de ADN difícil. En vista de estas limitaciones, actualmente se está realizando un estudio exploratorio de genotipado, mediante una *PCR* alelo-específica que posea un único paso de amplificación con *primers* marcados. De esta forma sería posible presentar un sistema alternativo que sea sencillo, más fiable y económico frente al *SNapShot*.

4. CONCLUSIONES



Conclusiones

Sobre el estudio genético de poblaciones americanas:

1. La obtención de nuevos datos genéticos en grupos nativo americanos, como los presentados en el análisis de SNPforID-52-plex, resultan de interés en el estudio forense y de poblaciones, debido a la necesidad de estimar frecuencias especialmente en grupos étnicos poco caracterizados.
2. El panel LACE de predicción del componente ancestral en poblaciones americanas, proporciona una estima óptima de las proporciones de mezcla individual, mediante el uso de AIMs informativos y equilibrados que presenten una medida de error equiparable entre los grupos ancestrales y un funcionamiento consistente frente a datos derivados del análisis de GW.
3. El panel LACE ha demostrado ser de gran utilidad para el control de la estratificación en estudios de asociación, empleando 314, 194, e incluso 88 AIMs. Por lo tanto, el uso de este panel contribuiría a minimizar el riesgo de falsos positivos en estudios de asociación con genes candidatos en poblaciones de América.
4. El empleo del panel LACE de inferencia de grupos ancestrales demostró la existencia de una fuerte concordancia entre la historia demográfica de cada población y las estimaciones de mezcla que fueron determinadas.
5. La estimación del tiempo desde que ocurrió el proceso de mezcla en algunas poblaciones de América, reveló en poblaciones afro-descendientes, un período que data entre 174-203 años, indicando que este proceso ha sido reciente en comparación con las poblaciones mestizas (mezcla Europea-Americano), cuya estimación ha sido entre 230-375 años, hechos que corresponden con los registros históricos conocidos.
6. El empleo de paneles de AIMs-SNPs, que incluyan la componente aportada recientemente por otros grupos continentales en América, además de los principales previamente descritos (Africa, Europa y Nativos), tales como PIMA en conjunto con 34plex, demostraron ser herramientas alternativas de utilidad en el estudio genético

de las poblaciones multi-étnicas que incluyan también asiáticos.

7. El panel de PIMA en conjunto con 34plex han demostrado con un bajo número de SNPs, aportar un poder de predicción de grupos ancestrales equiparable al panel LACE (de mayor número de marcadores), resultando de gran interés para su adaptación en estrategias de genotipado sencillas, lo que es de especial utilidad en el campo forense.

8. En cuanto al estudio del ADN mitocondrial, el componente nativo americano detectado es el que más prevaleció entre las poblaciones urbanas de Venezuela que fueron estudiadas. Además de la presencia de linajes nativos autóctonos, fue observada una contribución de otros perfiles nativos característicos de distintas regiones dentro del continente, reflejo de un flujo migratorio intra-continental.

9. A través del estudio de genomas completos mitocondriales en población urbana de Venezuela, se han descrito dos nuevos haplogrupos en el componente nativo: B2j y B2k, probablemente autóctonos. Adicionalmente, se han descrito otros nuevos sub-linajes mitocondriales en la localidad de Pueblo Llano, dos de ellos en A2 y uno en B2b, en los cuales se confirma una marcada deriva génica y consanguinidad documentada en esta región.

10. Existe una variabilidad significativa entre las frecuencias de haplogrupos mitocondriales observadas en las localidades urbanas de Caracas y Pueblo Llano. Estas diferencias coinciden con la historia demográfica documentada en ambas localidades, en las cuales se registra la presencia de un poblamiento originario de etnias nativas en Pueblo Llano, además de un contacto tardío de esta localidad con poblaciones europeas debido a su aislamiento, mientras que en Caracas los eventos de inmigración han estado ocurriendo desde la época colonial hasta nuestros días.

11. La mayoría de los haplogrupos mitocondriales L observados en Venezuela, proceden probablemente de África central y oeste, mientras que la proporción de haplogrupos europeos encontrados, proceden principalmente de España, Portugal e Italia, hechos que se ajustan casi perfectamente a los datos obtenidos a partir de los registros históricos existentes.

Sobre la predicción de EVCs en un contexto forense:

1. Se ha desarrollado un multiplex para el estudio de la pigmentación con propósitos forenses, el cual ha demostrado un funcionamiento óptimo en la predicción de color de ojos, tras su evaluación en diversas poblaciones europeas de gran variabilidad fenotípica.
2. El grado de información obtenido a partir de la selección de SNPs divergentes para la inferencia de pigmentación, ha sido tal que ha permitido realizar una predicción óptima del color de ojos azules y marrones, junto con una mejora en la verosimilitud de la predicción de fenotipos complejos (verdes-avellanas), en relación a otros estudios que han sido publicados.
3. El empleo de programas de clasificación bayesiano como STRUCTURE y *Snipper*, demostraron ser estrategias de utilidad en la selección de los conjuntos de entrenamientos estudiados, ofreciendo estimaciones sobre el error de clasificación y razón de verosimilitud, de gran valor en el campo forense y permitiendo definir a individuos de fenotipos complejos (verde-avellana), como un continuo que comprende un rango de variaciones de proporciones de mezcla genética y no como un cluster homogéneo.
4. A través de los análisis de asociación que fueron realizados en el presente trabajo de investigación, se confirma que los principales marcadores responsables de la predicción del color de ojos, particularmente de su tonalidad, se encuentran contenidos en los genes HERC2 y OCA2, seguido por los genes SLC45A2, SLC24A4, ASIP, TYR y TYRP.
5. Entre los SNPs más informativos en la predicción de color de ojos, determinado mediante el análisis de divergencia de Jensen y Shannon, está rs12913832 seguido por los marcadores rs1129038 y rs1667394 contenidos en HERC2. La mayor divergencia acumulada obtenida con estos marcadores fue observada tras la comparación de los fenotipos azul y marrón.

6. La evaluación realizada con las estimaciones de AUC, reveló un incremento significativo en la predicción de fenotipos azul y marrón, pero especialmente de verde-avellana, cuando fue incorporado al sistema *Irisplex* el marcador rs1129038 (HERC2). Este SNP demostró un incremento en los niveles de especificidad en las poblaciones azules y marrones, así como un salto significativo en la sensibilidad para grupos verde-avellana. Además, este marcador ha mostrado un efecto adicional en la predicción tras el ajuste con rs12913832 (HERC2).

7. El análisis de entropía realizado con la herramienta MDR, reveló la presencia de un efecto interactivo de tipo sinérgico entre los marcadores rs12913832 y rs1667394 en HERC2, cuando se compara el grupo verde-avellana con el resto. Dicha evaluación sugiere a rs1667394 como otro posible candidato que contribuye a la predicción de fenotipos verde-avellana.

8. Los multiplexes para la inferencia de pigmentación humana con aplicación en la predicción de color de ojos SHEP1 y SHEP2, han demostrado estar en la capacidad de combinar sensibilidad, especificidad, reproducibilidad y estabilidad, lo que se demostró tanto en estudios de poblaciones como tras su evaluación en muestras con un alto grado de inhibición y en estado de degradación, de interés forense.

Apéndice

Apéndice

Nuevos biomarcadores en genética forense

Maroñas O, Freire A, Santos C, Ruiz Y, Söchtig J, Lareu MV, Carracedo Á.

(*Patología y Biología Forense, Tomo II, Bosch SA. (2011):1151-1169 ISBN: 9788497908726*)

Nuevos biomarcadores en genética forense

Olalla MAROÑAS, Ana FREIRE, Carla SANTOS, Yarimar RUIZ,
Jens SÖCHTIG, M.V. LAREU y Ángel CARRACEDO

1. Polimorfismos nucleotídicos simples (SNPs)

En los últimos años han aparecido nuevos marcadores genéticos que son ya de utilidad para un cierto número de casos en los que los marcadores convencionales fallan y que han abierto además un nuevo abanico de aplicaciones del estudio del ADN humano con fines forenses.

Entre los nuevos polimorfismos de interés forense destacan los SNPs (*single nucleotide polymorphisms*) que son polimorfismos nucleotídicos simples muy abundantes en todo el genoma (existen más de 15 millones de SNPs conocidos) que se han ido caracterizando gracias a los diversos proyectos HapMap y los proyectos de resuenciación del genoma y los Indels son mutaciones o polimorfismos relacionados con la inserción o supresión de secuencias de ADN

Los SNPs tienen una serie de características que los hacen ideales para identificación humana. En primer lugar tienen tasas de mutación más bajas que los STRs, lo que les da una gran eficacia en investigaciones de parentesco. En segundo lugar el tamaño del producto amplificado puede ser muy pequeño, lo que les confiere ventajas para ADN degradado. Tercero existen numerosos métodos para su análisis incluyendo métodos de alto rendimiento, lo que les haría ideales para bases de datos aunque, al estar ya estas establecidas con STRs es difícil que las sustituyan.

Respecto al poder de discriminación, al ser polimorfismos habitualmente bialélicos, se necesitan más SNPs que STRs para lograr el mismo rendimiento. Aproximadamente 52 SNPs balanceados tienen el mismo poder de discriminación que los 15 STRs que se emplean en la mayoría de los multiplexes forenses.

El primer desarrollo de un conjunto de SNPs validado para fines forenses y actualmente el más popular es el desarrollado por el consorcio SNPfor ID

(<http://www.snpforid.org>) que eligió un grupo de 52 SNPs sin ligamiento entre ellos y altamente polimórficos en poblaciones humanas (Sanchez J.J., 2006). Existen frecuencias hoy día para la totalidad de los grandes grupos poblacionales y muchas poblaciones concretas y ha sido ampliamente validado para aplicaciones forenses (se puede conseguir un perfil con menos de 500 pg de ADN) (Musgrave-Brown *et al.*, 2007).

El método elegido para su análisis inicialmente fue una minisecuenciación mediante SNaPshot (Applied Biosystems) pero para fines forenses la tecnología de Genplex (una modificación de la tecnología de SNplex, ambas de Applied Biosystems que combina PCR y el uso de ligasas y usa modificadores de la movilidad electroforética) (Phillips *et al.*, 2007).

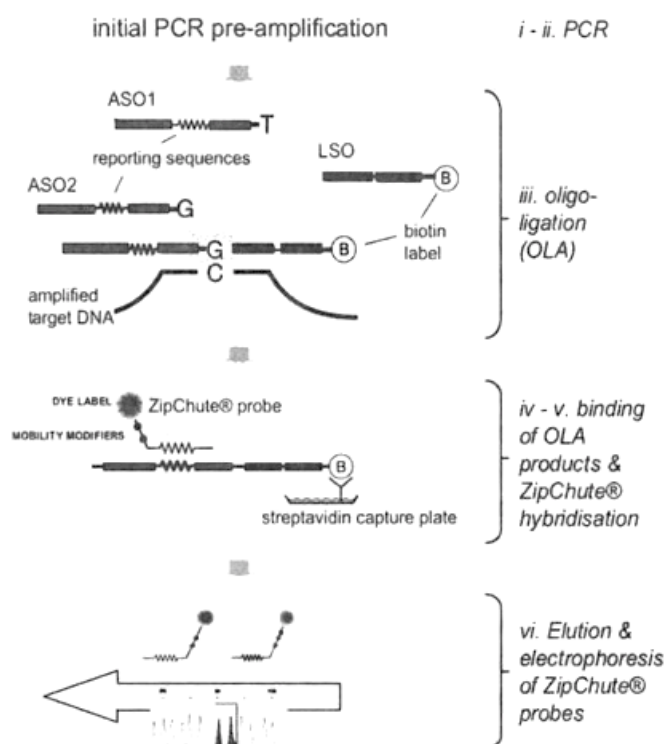


Figura 1. Principio del método de Genplex (de Phillips *et al.* 2007).

Existen numerosas tecnologías alternativas para el análisis de SNPs (Sobrino *et al.*, 2005) entre las que destaca el uso de espectrometría de masas MALDI-TOF combinada con minisecuenciación.

Quizá la aplicación más importante de los SNPs autosómicos es el análisis de muestras de ADN muy degradadas en las que los STRs e incluso los miniSTRs fallan. Se han descrito distintos trabajos donde se demuestra su eficacia

para este tipo de aplicaciones (Fondevila *et al.*, 2008) y también para analizar ADN con muy bajo número de copias, que es un problema, a fin de cuentas, de cuentas, relacionado con la degradación.

Se están desarrollando paneles de SNPs en regiones de ADN protegidas por nucleosomas (más resistentes a la degradación) y de SNPs trialélicos (pues la contaminación y mezcla de muestras es una de las limitaciones para el uso de los SNPs).

Los SNPs también están permitiendo solucionar casos complejos de parentesco (Phillips *et al.*, 2008) y el uso de chips de SNPs de alta densidad permite incluso solucionar relaciones lejanas de parentesco con eficacia.

Por último, además de SNPs autosómicos, se han desarrollado paneles de SNPs para ADN mitocondrial y cromosoma Y que nos permiten subdividir los haplogrupos de estos sistemas y aumentar la informatividad de las técnicas actuales. Es particularmente importante la subdivisión del haplogrupo H de ADN mitocondrial pues aumenta considerablemente la capacidad de discriminación de este último (Brandstätter *et al.*, 2006; Álvarez-Iglesias *et al.*, 2007).

2. Inserciones/deleciones (indels)

2.1. Origen y significado biológico

Los indels son inserciones o deleciones de una sola base, y en este sentido pueden ser incluidos en la categoría de los polimorfismos de base única (SNPs – Single Nucleotide Polymorphisms) (Jobling *et al.*, 2004), pero los eventos de inserción/delección ocurren menos frecuentemente que los fenómenos de sustitución (Zhang and Gerstein, 2003). En general los SNPs son diez veces más frecuentes que los indels pero esta diferencia es relativa y puede variar entre loci dependiendo de la secuencia contexto (Ebersberger *et al.*, 2002 in Jobling *et al.*, 2004). Por ejemplo, se sabe que el número de indels que se puede encontrar en regiones con función conocida (en gran parte ADN codificante) o afectadas por fuerzas evolutivas es muy reducido (Clark *et al.*, 2007). Por otra parte, la tasa de mutación de los indels cortos (por loci por generación) es muy similar a la de los SNPs – $2,3 \times 10^{-9}$ (Nachman y Crowell 2000 in Jobling *et al.*, 2004). Esta baja tasa de mutación indica que los indels presentan identidad por ascendencia, es decir, la presencia de una determinada base polimórfica en dos alelos implica que ambos han heredado esa base de un ancestro común (Jobling *et al.*, 2004).

Se sabe que los indels son abundantes en el genoma de organismos modelo. Así se espera que lo mismo ocurra en el genoma humano. Un estudio de Dawson *et al.* (2001) ha sugerido que los indels representan 18% de los poli-

morfismos del cromosoma 22 humano, con lo que se podría inferir que estos representan entre el 16% y el 25% de todos los polimorfismos en el genoma humano (Mills *et al.*, 2006).

Los indels dialélicos más comunes son los que presentan una diferencia de pocos nucleótidos entre sus alelos (Weber *et al.*, 2002). Nachman y Crowell (2002) han calculado que los indels cortos (1-20bp) se corresponden con el 10% de las modificaciones de secuencias no codificantes del genoma humano (Jobling *et al.*, 2004). Por otro lado, en el trabajo de Bhangale *et al.* (2005) se calculó a partir de la extrapolación de los datos de un conjunto de genes representativos de todo el genoma que, aproximadamente, uno de cada 15 polimorfismos dialélicos en las regiones intergénicas del genoma humano corresponden a un indel. De acuerdo con Clark *et al.* (2007) éstos pueden ocurrir a una tasa de 15 inserciones/deleciones por cada 100Kbp (las cuales corresponden aproximadamente a 43bp por cada 100Kb del genoma).

Otras características de los indels son la menor tasa de ocurrencia de supresiones comparada con la de inserciones y la preferencia por diferencia reducida entre la longitud de los dos alelos. Un estudio de Zhang y Gerstein (2003) ha demostrado que la longitud media de las inserciones y supresiones es de aproximadamente 4bp con un mayor número de observaciones para una longitud de 2bp.

Los indels, en conjunto con las sustituciones nucleotídicas y las reorganizaciones genómicas, son uno de los principales mecanismos inductores de mutaciones y defectos génicos, siendo por eso muy importantes para la variabilidad genética y evolución molecular (Britten *et al.*, 2003; Makova *et al.*, 2004; de la Chaux *et al.*, 2007), pudiendo tener un papel importante en el establecimiento y manutención de la diversidad del genoma humano (Robledo *et al.*, 2003).

Esta clase de polimorfismos apenas puede ser reconocida cuando se comparan múltiples secuencias homólogas de ADN y se detectan diferencias en la longitud total de esas secuencias (Pearce, 2006). Weber *et al.* (2002) estudió cerca de 2000 indels y concluyó que la gran mayoría tiene origen posterior a la divergencia de los antepasados comunes entre humanos, chimpancés y gorilas, pero la tasa de mutación de los indels es similar en los cromosomas X e Y de los primates (Sundstrom *et al.*, 2003). Esta observación indica que no existe sesgo relacionado con el sexo y, por tanto, que la mutación de estos marcadores puede ser el resultado de la acción combinada de la recombinación y número de divisiones celulares (Sundstrom *et al.*, 2003). Los errores introducidos durante la biosíntesis del ADN están correlacionados con las mutaciones puntuales y sus efectos biológicos descritos (García-Díaz and Kunkel, 2006). En el estudio de Makova *et al.* (2004) fueron encontradas evidencias de que indels largos y cortos pueden tener origen en mecanismos moleculares distintos: los primeros relacionados con la recombinación y los segundos con la

replicación. En el mismo estudio dichas conclusiones no fueron observadas en roedores, siendo éste el primer ejemplo reportado de sesgo relacionado con el sexo en mamíferos, y, de acuerdo con los autores, supone una evidencia que apoya la teoría del origen de los indels en la replicación celular (Makova *et al.*, 2004), pero existen otros mecanismos de mutación que afectan a los dos sexos de la misma forma, por ejemplo los daños en el ADN (Sundstrom *et al.*, 2003). Otro factor a tener en cuenta es la localización del indel; los que ocurren en regiones codificantes van a estar sometidos a presiones selectivas más fuertes debido al impacto que pueden tener en la estructura de la secuencia de aminoácidos o en las propiedades de los genes en los que se localizan (Taylor *et al.*, 2004; de la Chaux *et al.*, 2007).

En el genoma existen «hotspots» de indels, es decir, regiones que tienen tendencia a presentar un número elevado de indels en comparación con el resto del genoma, que pueden corresponder a «hotspots» de recombinación (Mills *et al.*, 2006; Costantini and Bernardi, 2009). Muchas de estas regiones presentan también un elevado número de SNPs, y por lo tanto, pueden ser consideradas como regiones de variación genética (Mills *et al.*, 2006). Así, hay autores que sugieren que los indels aumentan la tasa de sustituciones que ocurren en las regiones a su alrededor (Tian *et al.*, 2008; Chen *et al.*, 2009). En general, la correlación indels-SNPs es considerada indirecta ya que sería una respuesta a la composición o estructura de la secuencia o incluso a una restricción funcional de esa región (Tian *et al.*, 2008). Otra hipótesis sugerida y confirmada en el estudio de Tian *et al.* (2008) es la mutación inducida por los indels, es decir, estos (en estado heterocigótico) aumentan la sustitución de nucleótidos en posiciones adyacentes.

Una conclusión muy interesante obtenida en el estudio de Costantini y Bernardi (2009) es la relación que se observa entre el número elevado de inserciones y supresiones que ocurren en «isochores» – grandes regiones de ADN ricas en contenido GC. Por otra parte, los SNPs presentan una distribución uniforme a lo largo de estas regiones, lo que puede ser explicado por el origen de estos polimorfismos: errores durante la replicación del ADN los cuales no son afectados por la secuencia contexto (Costantini and Bernardi, 2009).

2.2. Aplicaciones de los indels

Una vez mencionado que los indels se corresponden con una gran parte de los polimorfismos dialélicos humanos, su inclusión en los estudios puede permitir una mayor resolución de los mapas genéticos y una mejora de la estimación de las tasas de recombinación y de los haplotipos inferidos a través de métodos estadísticos (Bhangale *et al.*, 2005).

Weber *et al.* (2002) recomiendan la utilización de indels para la mayoría de los estudios genéticos. Como el número de estudios sobre la distribución de indels a un nivel intraespecífico es reducido (Pearce, 2006), hay que tener cuidado de elegir conjuntos de marcadores con aproximadamente el mismo nivel de información para cada grupo poblacional, de forma que no se cometan errores en la estima de la diferenciación poblacional (Weber *et al.*, 2002; Pearce, 2006).

La utilización de técnicas de genotipado en casos criminales es una forma de establecer la identidad de evidencias derivadas de cualquier material biológico. Se ha observado un gran desarrollo en el área forense, lo que permitió el aumento de la resolución y sensibilidad de la detección (Budowle and van Daal, 2008). Pero a pesar de todos los avances observados hay muestras de ADN que no son susceptibles de análisis con los marcadores validados en forense, los STRs. Esto puede deberse a la baja concentración de ADN o el estado de degradación de la muestra (Budowle and van Daal, 2008). En estos casos la utilización de SNPs es una alternativa viable ya que estos marcadores presentan un conjunto de características muy favorables. Por un lado permiten la amplificación de muestras degradadas con una mayor tasa de éxito ya que el polimorfismo está constituido por una sola base y apenas se necesita un pequeño fragmento intacto de ADN (60–80bp), por lo tanto la probabilidad de encontrar esa secuencia es mayor (Gill *et al.*, 2004; Phillips *et al.*, 2004; Budowle and van Daal, 2008). Otra ventaja es su estructura simple sin secuencias repetitivas y la posibilidad de desarrollar reacciones multiplex que incluyen un gran número de marcadores y que pueden ser analizadas a través de tecnologías de elevada rentabilidad, lo que permite beneficiarse al máximo de las ventajas de los SNPs utilizando una cantidad mínima de muestra (Gill, 2001; Gill *et al.*, 2004; Phillips *et al.*, 2004; Sobrino *et al.*, 2005; Budowle and van Daal, 2008). Ésta última es una de sus características con mayor interés en el área de la genética forense, aunque también resulta muy interesante la baja tasa de mutación (10^{-8}) (Nachman y Crowell 2000 en Jobling *et al.*, 2004), la cual permite que estos sean marcadores muy estables y por lo tanto adecuados para análisis basados en linajes o cuando no existe una muestra directa de referencia (Budowle and van Daal, 2008). La mayoría de los SNPs son bialélicos (y por lo tanto individualmente no son tan informativos como los loci STR seleccionados para forense), por lo que es necesario analizar un mayor número de SNPs (50–100) para alcanzar el mismo nivel de discriminación obtenido con el análisis de STRs (Gill, 2001; Gill *et al.*, 2004; Phillips *et al.*, 2004; Budowle and van Daal, 2008). Otras limitaciones de la utilización de SNPs son la dificultad en la interpretación de mezclas ya que son marcadores binarios, así como la reducción del tamaño del amplicón pues la probabilidad de contaminación es mayor (Phillips *et al.*,

2004; Budowle and van Daal, 2008). Estos dos problemas pueden ser fácilmente solucionados a través de la utilización de SNPs no-binarios, es decir, los que indican la presencia de ADN exógeno a través de la detección de un tercero o cuarto alelo (Phillips *et al.*, 2004).

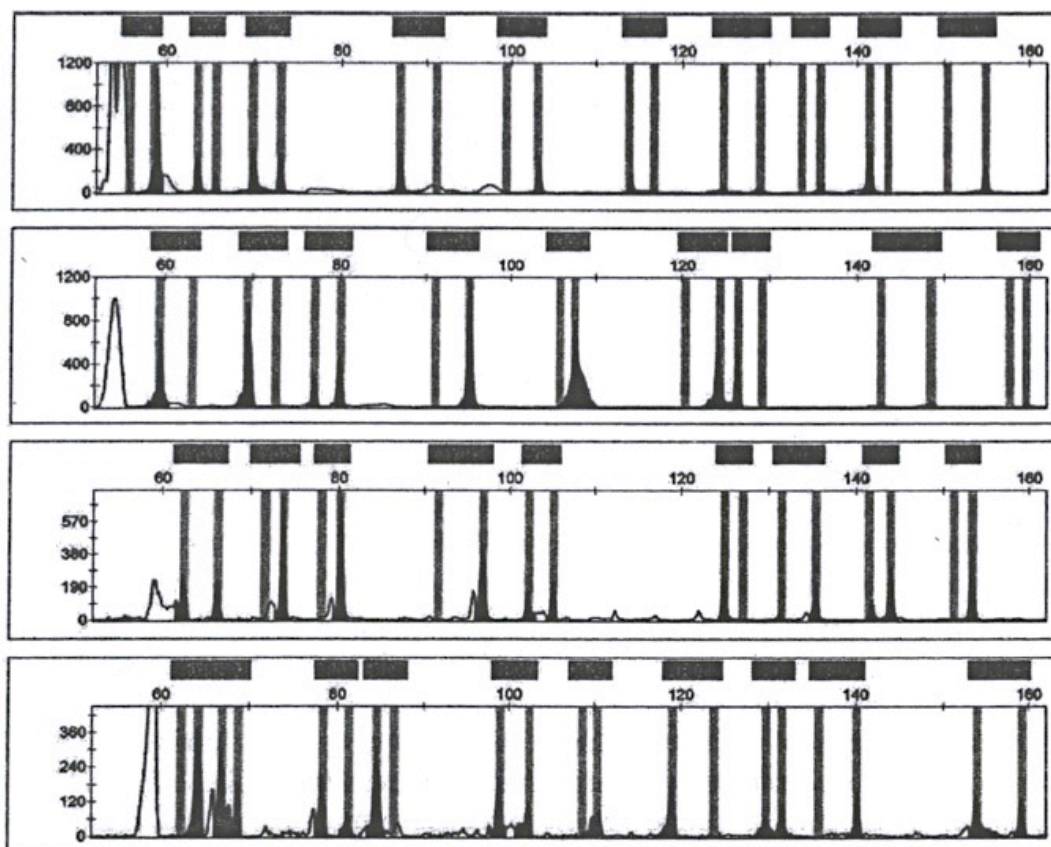


Figura 2. Electroferograma del e8 indelplex de Pereira *et al.* (2009).

Los indels son marcadores con un gran potencial en el área de la genética forense ya que combinan las características de los SNPs y de los STRs (Pereira *et al.*, 2009). Tal y como se ha mencionado anteriormente, los indels son polimorfismos binarios, con lo cual, comparten las mismas características que los SNPs y por lo tanto tienen las mismas ventajas en genética forense (Edelmann *et al.*, 2009). El grupo más numeroso de indels dialélicos, y que está ampliamente distribuido en el genoma, es el que contiene polimorfismos con una diferencia entre alelos de 3-15 nucleótidos (Mills *et al.*, 2006; Pereira *et al.*, 2009; Santos *et al.*, 2010) y por lo tanto pueden ser analizados con técnicas establecidas para STRs (por ejemplo, amplificación de fragmentos utilizando primers marcados con colorantes fluorescentes seguida de detección a tra-

vés de electroforesis capilar) pero con reducido tamaño de amplicón, lo que mejora la amplificación de muestras degradadas y la creación de reacciones multiplex (Edelmann *et al.*, 2009; Pereira *et al.*, 2009; Ribeiro-Rodrigues *et al.*, 2009; Pimenta and Pena, 2010; Santos *et al.*, 2010). Además, presentan frecuencias distintas entre poblaciones (Hofer *et al.*, 2009; Ribeiro-Rodrigues *et al.*, 2009) y por lo tanto tienen potencial de AIMs (Ancestry Informative Markers) (Weber *et al.*, 2002; Pereira *et al.*, 2009; Santos *et al.*, 2010) y su inclusión en el análisis puede influir positivamente en la estima de la diferenciación entre poblaciones (Pearce, 2006) lo que es importante, por ejemplo, para detectar subestructura poblacional en estudios de casos-controles (Santos *et al.*, 2010).

Varios estudios fueron ya realizados utilizando indels para identificación humana (Pereira *et al.*, 2009; Li *et al.*, 2011), pruebas de parentesco (Edelmann *et al.*, 2009; Pimenta and Pena, 2010) y caracterización de poblaciones humanas y diferenciación entre éstas (Bastos-Rodrigues *et al.*, 2006; Ribeiro-Rodrigues *et al.*, 2009; Santos *et al.*, 2010).

Con esto, la principal conclusión es la utilidad de conjuntos de 40-50 indels bien seleccionados para la detección de eventos muy antiguos en la historia demográfica de las poblaciones (Bastos-Rodrigues *et al.*, 2006; Santos *et al.*, 2010), incluyendo la mezcla entre poblaciones y el cálculo de proporciones de ancestralidad (Ribeiro-Rodrigues *et al.*, 2009; Santos *et al.*, 2010), a través de un análisis simple y que utiliza técnicas ya disponibles en los laboratorios de genética (Ribeiro-Rodrigues *et al.*, 2009). Además, la utilización de indels altamente polimórficos en determinadas poblaciones y bien distribuidos en el genoma permite el desarrollo de conjuntos de marcadores útiles en la identificación individual y en casos de parentesco, incluso en muestras muy degradadas (Pereira *et al.*, 2009; Pimenta and Pena, 2010).

3. Marcadores de ancestralidad (AIMS)

El término ancestralidad, como su propio nombre indica hace referencia a nuestros ancestros, a nuestros antepasados, a nuestras raíces. Cada ser humano es el resultado de la recombinación entre dos moléculas de ADN, cada cual a su vez, ha sido producto de múltiples recombinaciones anteriores a lo largo de los siglos de la humanidad, desde que la primera población humana surgió, hace 100.000-200.000 años en un lugar recóndito del Este de África. Análisis de cromosoma Y, ADN mitocondrial y marcadores autosómicos, han revelado una estructura geográfica de las poblaciones humanas a nivel continental, sugiriendo a su vez, la migración de un pequeño grupo de individuos procedente del Este de África (out of Africa) y la consecuente expansión de sus

descendientes para resultar en las poblaciones actuales que habitan hoy día la Tierra (Lewin, 1987).

Generalmente, las poblaciones que presentan un pequeño tamaño efectivo de la población y/o han experimentado un marcado cuello de botella, tendrán un pronunciado flujo genético, resultando en un rápido aumento en frecuencias alélicas derivadas. En cambio, poblaciones que han mantenido un amplio tamaño o han experimentado expansión, tenderán a preservar variantes ancestrales.

La porción no codificante del ADN es una herramienta tan poderosa que, nos permite detectar variantes alélicas asociadas a determinadas poblaciones. Esto supone una importantísima ventaja a la hora de inferir ancestralidad biogeográfica en una muestra. Esta posibilidad se convierte en una aplicación clave en el *campo forense*. Ante un crimen o delito, no siempre es posible cotejar una muestra dubitada presente en la escena del crimen, con una muestra de referencia, es decir, con una muestra de un sospechoso. Sin embargo, en caso de lograr inferir la ancestralidad de una determinada muestra, sería posible pues, aportar un dato relevante a la investigación, puesto que no es lo mismo no poseer dato alguno a poder conocer el origen geográfico asociado a la muestra dubitada, esto delimita el radio de acción y por tanto incrementa las posibilidades de éxito en la investigación policial evitando que ciertos casos se cierren por falta de pruebas.

Los marcadores genéticos que nos otorgan esta posibilidad son los llamados AIMs (*ancestral informative markers*). Los AIMs son polimorfismos cuyo genotipo infiere información ancestral en muestras/individuos. Estos marcadores genéticos se caracterizan por sus diferencias en las frecuencias alélicas entre diversas poblaciones; por ejemplo, un AIM bialélico puede presentar en la población 1, una frecuencia alélica de 0.9 para el alelo A y una frecuencia alélica de 0.1 para el alelo B mientras que en la población 2 la situación es inversa, de tal modo que, el alelo A presenta una frecuencia alélica de 0.1 en tanto que el alelo B presenta una frecuencia alélica de 0.9. Esto nos permite inferir que, en caso de observar un alelo A en una muestra analizada en el laboratorio, la probabilidad de que esa muestra pertenezca a la población 1 es considerablemente mayor de que pertenezca a la población 2. Es obvio que un único marcador no permite obtener exactitud en el resultado dado, que aún se tendría presente un 10% de error. Esta situación es solventable mediante el incremento del número de marcadores, así el error disminuye hasta llegar al punto en el que se puede obtener con un elevado poder de discriminación, un dato que permitirá inferir la ancestralidad de esa muestra dubitada. Es importante tener en cuenta que estos marcadores nos revelan el origen ancestral de la muestra pero no infieren rasgos fenotípicos de la misma, si bien, indirectamente, se puede establecer un fenotipo en base a la correlación de la

expresión fenotípica con ciertos elementos de la estructura ancestral de las poblaciones humanas.

Diferentes clases de polimorfismos son candidatos para actuar como AIMs, existiendo actualmente grandes bases de datos de acceso libre donde es posible encontrar toda la información necesaria respecto a estos marcadores (NCBI, USCS Genome Browser, Ensemble...). Los SNPs, polimorfismos de secuencia, son hasta el momento los mejores marcadores ancestrales debido a su estabilidad (baja tasa de mutación), su densidad de distribución a lo largo del genoma y a su amplio rango de patrones de frecuencias alélicas que presentan entre diversas poblaciones. Otros marcadores genéticos también pueden ser estudiados como AIMs, pero presentan ciertas desventajas respecto a los SNPs. Los microsatélites, por ejemplo, la batería de marcadores más importante en el análisis forense, debido a su elevado poder de discriminación, se vuelve en cierto modo carente de valor en este contexto debido a que no presentan grandes contrastes en las frecuencias alélicas entre poblaciones para poder ser analizados como un panel de menos de 50 loci, sino que presentan muchos alelos compartidos entre poblaciones (Budowle and van Daal, 2008). Por otra parte están el ADN mitocondrial y el cromosoma Y que proporcionan información filogeográfica pero su característica haploide hace necesarias amplias bases de datos para medir adecuadamente la variabilidad poblacional. Otra clase de polimorfismos candidatos en la inferencia de ancestralidad son los Indels (Insertion Deletion Polymorphisms). Collins-Schramm *et al.* ya han estudiado la subestructura de una población de mejicanos americanos mediante un set de 35 indels (Collins-Schramm *et al.*, 2004).

Independientemente de la clase polimorfismo, los AIMs se pueden clasificar a su vez, en cuatro grupos:

- Marcadores específicos de población: loci con un polimorfismo detectado en uno o dos grupos poblacionales pero ausente en los demás.
- Marcadores con frecuencias alélicas sesgadas: loci con un alelo común en una población, que es raro en las demás ($<0,6$).
- Marcadores trialélicos: presentan múltiples sustituciones en la misma posición.
- Marcadores fijados: loci en los que un alelo se observa exclusivamente en un grupo poblacional y el alelo alternativo, exclusivamente en la otra.

Diversos paneles de AIM-SNPs han sido publicados recientemente. El primer test forense capaz de inferir origen geográfico dentro los tres grandes grupos poblacionales: Africano sub-Sahariano, Europeo y Asiático ha sido

desarrollado por Phillips *et al.* (2007) teniendo su principal aplicación en la investigación policial del atentado en Madrid del 11-M (Phillips *et al.*, 2009). A continuación se han desarrollado otros paneles en los que se aumenta el número de AIM-SNPs analizados y también el número de poblaciones capaces de diferenciar (Halder *et al.*, 2008; Nassir *et al.*, 2009). Para poblaciones europeas, ha surgido un panel capaz de diferenciar un gradiente norte-sur así como un gradiente este-oeste (Tian *et al.*, 2008).

Rosenberg *et al.* (2002) fueron los primeros en estudiar la estructura de las poblaciones humanas. Para ello emplearon 377 microsatélites autosómicos en 1056 individuos procedentes del Human Genome Diversity Panel [HGDP-CEPH panel pertenecientes a 52 poblaciones. Se identificaron seis principales clusters genéticos, cinco de ellos, correspondientes a los cinco grandes grupos continentales (África, Este Asiático, Oceanía, América y Euroasia, pudiendo ser ésta dividida en: Europa, Medio Oriente y Asia Sur y Central) correspondientes a su vez a la mayores barreras físicas (océanos, Himalaya y Sáhara). El sexto componente correspondía a una población pequeña y aislada, el grupo Kalash, cuyo lenguaje es indo-europeo y habita en el norte de Pakistán.

Si bien es cierto que existen individuos con una ancestralidad muy pura, dado que sus antepasados no han intercambiado material genético con individuos pertenecientes a otras etnias, también hemos de tener en cuenta la existencia de otros individuos que presentan un importante grado de mezcla, es decir, su componente genético ancestral no es posible asignarlo en el 100% a un único grupo poblacional, sino que está compuesto por porcentajes diferentes relativos a diversos grupos poblacionales. Un ejemplo de esta situación se encuentra reflejado en países de América del Sur (Kosoy *et al.*, 2009). Los primeros habitantes de estas tierras, indígenas nativoamericanos, con una variabilidad genética específica, vieron disminuido su componente genético nativoamericano al ser combinarse su carga genética con otros grupos poblacionales, por una parte, europeos procedentes de la conquista de América y por otra parte, africanos procedentes del tráfico de esclavos. Estos dos acontecimientos, que implicaron una masiva invasión poblacional, modificaron el curso de la historia así como la variabilidad genética de estas poblaciones, existiendo actualmente un importante grado de mezcla en poblaciones localizadas en estas zonas geográficas. Aún así, el componente nativoamericano ha logrado sobrevivir y continúa estando presente.

Un punto muy importante a resaltar es la importancia que toma la detección de la ancestralidad a la hora de desarrollar estudios de asociación (Tian *et al.*, 2008) cada vez más comunes en estos tiempos. En este tipo de estudios, casos y controles son analizados para buscar posibles asociaciones entre una variante genética y un determinado fenotipo. En los últimos años muchas

variantes alélicas han sido asociadas a una gran variedad de enfermedades complejas gracias al desarrollo de los estudios GWAs (*Genome-Wide Association*) (Bottini *et al.*, 2004). Debido a la complejidad en la interpretación de los resultados, el análisis de estos datos ha de ser muy cuidadoso para evitar falsas asociaciones. Una variante genética puede estar asociada con una determinada enfermedad; imaginemos que esta misma variante genética sólo se encuentra presente en la población 1 pero no en la población 2. Si los casos perteneciesen a la población 1 y los controles a la población 2, en caso de encontrar asociación para la variante alélica X, estaríamos ante un caso de falso positivo. Es primordial conocer que la subestructura de los casos y controles se encuentra balanceada y es similar entre ambas poblaciones para así poder corregir el error causado por estratificación. Esto es de especial relevancia en poblaciones en la que ya se espera a priori un elevado grado de admixture.

Existen cuatro causas principales de falsos positivos: umbrales estadísticos inapropiados, artefactos genotípicos, factores medioambientales o diferencias ancestrales no reconocidas causadas a su vez, por errores de genotipado y aplicación de umbrales estadísticos no adecuados. Ante esta situación se han desarrollado métodos para corregir estos errores (Price *et al.*, 2006; Epstein *et al.*, 2007).

Los paneles de AIM-SNPs, pueden ser empleados no sólo para determinar subestructura en muestras de casos y controles ya analizados, sino también, como método de cribado inicial para seleccionar las muestras que pueden ser candidatas a un análisis mediante GWAs y así evitar la pérdida de información causada por posibles falsos positivos.

Cada día nuevos marcadores y nuevas poblaciones son estudiados. Tener una amplia batería de AIMS puede parecer en principio que aumenta en gran medida el poder de discriminación, pero el genotipado a gran escala no está al alcance de todos los laboratorios debido a su elevado coste. Nuevos estudios comienzan a centrarse en la idea de que un número más reducido de marcadores puede proporcionar la misma información versus un elevado set de los mismos (Ruiz-Narváez *et al.*, 2011).

A parte de todo lo comentado hasta el momento hay un punto final sobre el debemos reflexionar. Las circunstancias actuales favorecen un flujo continuo de individuos de unos países a otros. Las mezclas entre personas de diferentes poblaciones es cada poco más palpable de tal modo que, si pensamos en generaciones futuras, aunque muy lejanas, podremos observar como nuevas mezclas genéticas aparecen, como grupos poblacionales puros disminuyen su tamaño efectivo, como un nuevo concepto de diversidad poblacional surge, pero para ello, aún deben pasar muchos años venideros y para entonces nuevas estrategias que ayuden a comprender y esclarecer estos nuevos retos se habrán desarrollado.

4. Pigmentación humana

En Genética Forense la predicción de la pigmentación humana como rasgo físico ofrece grandes ventajas para dirigir la investigación criminal cuando no hay un perfil genético de referencia con que comparar, reduciéndose así el universo de sospechosos en la población.

En el supuesto caso de un individuo perteneciente a una población mezclada, la contribución mayoritaria de marcadores de ancestralidad de un determinado grupo étnico a dicha mezcla, puede predisponernos a inferir una apariencia física errónea, y es en este caso particular, en donde el análisis de las características físicas como la pigmentación, adquiere vital importancia. Pero además, el conocer el color de los ojos, piel y cabello podría ayudar a evitar errores en la identificación de posibles sospechosos, dado que, a consecuencia del trauma, un alto porcentaje de víctimas de un delito se equivocan en las ruedas de reconocimiento.

En cuanto a las técnicas de diagnóstico empleadas, considerando que, en la escena de un crimen se suele encontrar poca cantidad de muestra o incluso ADN degradado, se hace importante el uso de técnicas sensibles que permitan detectar varios marcadores de forma conjunta. Una buena estrategia es el uso de multiplexes de SNPs (polimorfismos de una única base) que es necesario que tengan unas frecuencias alélicas que varíe entre distintas poblaciones para que sean informativos (Phillips *et al.*, 2007). Técnicamente también es fundamental un modelo de clasificación fenotípica apropiado para las distintas poblaciones, y el diseño de un test genético sustentado en un análisis estadístico, que provea un grado de error asociado a la probabilidad de presentar un tipo específico de color.

4.1. Aspectos generales

La pigmentación al igual que otras características físicas humanas, es un carácter complejo que posee un fuerte componente genético y que es el resultado de la acción de múltiples genes, así como de factores ambientales, eventos estocásticos y epigenéticos (Jobling *et al.*, 2004).

Es un rasgo físico altamente heredable y variable entre los seres humanos (Jablonski and Chaplin, 2000; Rees, 2004). Dicha variación se debe mayoritariamente a diferencias en la cantidad, tipo y distribución de la melanina, que es producida en los melanocitos durante la melanogénesis (Sturm, 1998; Sturm, 2006; Branicki *et al.*, 2009). En esta ruta se producen dos tipos de pigmentos, eumelanina que dan lugar a pigmentos oscuros (marrón oscuro y negro), y feomelanina que dan lugar a pigmentos claros (rojo y amarillo) (Minwalla *et al.*, 2001; Barsh, 2003). La variedad intra e interpoblacional en cuanto a la

pigmentación depende no solo del número, tamaño y distribución de los melanosomas (pues el número de melanocitos es, aproximadamente, el mismo en todos los individuos) (Sulem *et al.*, 2007), sino también de la combinación de estos pigmentos en distintos ratios (Sturm, 1998).

4.2. Pigmentación de ojos

El color de ojos como rasgo físico a inferir en la investigación forense precisa un estudio de la morfología, bioquímica y genética del iris. Esta estructura forma parte de la capa anterior del tracto uveal del ojo (Forrester, 2008), y se encuentra constituida por dos capas de tejidos: el epitelio pigmentado del iris (IPE) y el estroma iridial. Ambas capas contienen melanocitos pero es el estroma quien contribuye mayormente a la coloración del ojo (Imesch *et al.*, 1997). En la actualidad hay un consenso en que las variaciones de tonalidades del color de iris pueden ser atribuidos principalmente a la variabilidad, número y distribución de melanocitos en el estroma, sin embargo, resulta de interés considerar que la presencia de «patrones» en el iris contribuyen también a la impresión visual de la coloración del ojo, tales como las criptas de Fuchs, surcos de contracción, anillo peripupilar, cúmulos de colágeno, etc. Éstos se encuentran bajo un fuerte control genético, y actualmente, se han descrito sus posiciones en el genoma, e incluso, en algunos casos los genes involucrados (Sturm, 2004; Sturm, 2009).

En la coloración del ojo, los melanocitos con altos niveles de melanina absorben más luz, dando la apariencia de coloración oscura. Cuando hay una ausencia de melanina en la capa anterior del ojo, la luz entra por el estroma y es dispersado por el colágeno el cual absorbe la mayoría de los colores excepto el azul y el gris. Los tonos intermedios como el verde y avellana resultan de diversas cantidades de melanina que permiten que la luz entre a través del estroma reflejando una mezcla de tonalidades (Imesch *et al.*, 1997). Sin embargo, ha de reconocerse que esta clasificación de grupos fenotipos resulta una simplificación, sabiendo que existe un amplio rango de tonalidades en el iris (Sturm, 2004).

Uno de los locus responsables del fenotipo color azul fue primero identificado por estudios de ligamiento en el cromosoma 15q por Eiberg y Mohr en 1996 (OMIM 227220). El locus candidato fue analizado a través de mapa genético en la región 15q12-13, y posteriormente, OCA2 ubicado dentro de esta región fue sugerido como gen candidato. El transcrito del gen OCA2 produce variaciones en la proteína integral de membrana «P» que ayuda a la regulación de la melanogénesis. Se ha sugerido que la función de OCA2 es la de ser un antiporte Na^+/H^+ (Eiberg *et al.*, 2008) o un transportador de glutamato (Bennett, 2003), funciones que indican que OCA2 está involucrado

en el tráfico intracelular de la enzima tirosinasa durante la maduración del melanosoma (Toyofuku *et al.*, 2002). Por su parte el gen HERC2 ubicado en la región 5'UTR de OCA2 fue descrito posteriormente como asociado significativamente al color de ojos azul y marrón, también por estudios de ligamiento y asociación (Sulem *et al.*, 2007; Han *et al.*, 2008; Mengel-From *et al.*, 2009; Eriksson *et al.*, 2010). Para los SNPs rs12913832 y rs1129038 contenidos en esta región, se ha sugerido una actividad reguladora sobre el promotor de OCA2 (Eiberg *et al.*, 2008), descrito posteriormente en un modelo presentado por Sturm (2009). En dicho modelo la presencia del alelo T para el SNP rs12913832 induce el desempaqueado de la heterocromatina en esta región HERC2-OCA2, lo cual a su vez permite que el promotor de OCA2 esté disponible para la transcripción. Consecuentemente, el factor HLTF puede actuar como secuencia específica de reconocimiento, para que se unan los factores de transcripción MITF y LEF1 específicamente sobre el locus de la región control, induciendo entonces la transcripción de la proteína OCA2 que estimula la maduración del melanosoma. El resultado es la aparición del color marrón. Por otra parte un cambio de base en rs12913832 previene la interacción del factor HLTF con la heterocromatina, evitando la unión de los factores MITF y LEF1, y por consiguiente la región promotora de OCA2 permanece cerrada, la ausencia de su proteína entonces conlleva a la formación de melanocitos inmaduros resultando en la aparición del color azul (Sturm, 2009).

Recientemente, han sido propuestos modelos de predicción empleando el genotipado de estos SNPs, así como en otros genes de menor impacto como SLC45A2, TYRP1, SLC45A4 e IRF4. Dichos modelos están basados en el análisis probabilístico de regresión logística multinominal contribuyendo así a la inferencia de fenotipos azules y marrones de manera significativa, y con una sensibilidad menor en fenotipos intermedios (Liu *et al.*, 2009). De igual manera se han propuesto métodos de cuantificación del color del iris basados en el análisis de fotografía digital, estimando valores de saturación y tonalidad que buscan aportar más información en la detección de genes asociados al color de ojos (Liu *et al.*, 2010). Sin embargo, en la actualidad aún continúa la investigación para el diseño de pruebas genéticas que puedan proveer mejoras en la inferencia fenotípica, tomando en cuenta las limitaciones que presenta este complejo rasgo físico.

4.3. Pigmentación del cabello

Aunque la pigmentación del cabello está distribuida como un rasgo continuo, suele categorizarse en cuatro grandes grupos: pelirrojo, rubio, marrón y negro. Esta variabilidad se observa principalmente en poblaciones con ascen-

dencia europea, presentando su mayor diversidad en la región geográfica del Báltico, comprendiendo Europa del Este y del Norte (Harrison, 1973; Frost, 2006). Cada uno de estos grupos de color de cabello se caracteriza generalmente por la cantidad, distribución, tamaño y forma de los melanosomas, así como su depósito de melanina (Ortonne, 1993; Liu *et al.*, 2005). El factor decisivo es el tipo dominante de pigmento y la proporción de mezcla de los pigmentos eu- y feomelanina. Hay proporciones individualmente muy diferentes (Borges *et al.*, 2001): pelirrojo y rubio contienen en total menos melanina que el marrón y negro. Para el cabello rojo, la proporción de eumelanina es del 67% y la de feomelanina del 33%. El rubio posee un 95% de eumelanina igual que el cabello marrón; estos comparten con el negro la misma cantidad de melanosomas, con la diferencia de que los del rubio son más pequeños y redondeados y los del marrón de forma elipsoidal. Finalmente el cabello negro muestra los melanosomas más grandes y más densos (99% eumelanina).

Desde el punto de vista genético forense, se han identificado en los últimos años numerosas regiones cromosómicas y variantes genéticas (principalmente SNPs), las cuales muestran una asociación con el color del cabello (Sturm, 2008). En la actualidad, sólo se puede predecir con alta precisión el color pelirrojo (Grimes *et al.*, 2001). El responsable de este fenotipo es el gen del receptor de melanocortina (MC1R, MIM 155555). Este codifica una proteína integral de membrana que es un punto de control clave en la melanogénesis. Las mutaciones que generan una pérdida de función en la proteína MC1R favorecen la producción de feomelanina. La mayoría de las personas pelirrojas son heterocigotos compuestos u homocigotos para las mutaciones R151C, R160W, D294H, R142H, D48E y comparten otros fenotipos comunes, tales como piel clara y pecas (Rees, 2003). Además, recientemente se encontraron SNPs de otros genes asociados con colores del cabello no pelirrojo (Sulem *et al.*, 2007; Han *et al.*, 2008; Sulem *et al.*, 2008; Eriksson *et al.*, 2010), para los cuales se han hecho predicciones usando 13 SNPs en 11 genes. Estos estudios revelaron que el cabello negro se podría inferir correctamente en un 87% de los casos, mientras que marrón y rubio en más de un 80% de los casos en población polaca (Branicki *et al.*, 2011).

4.4. Pigmentación de piel

El color de la piel, al igual que el cabello, es un rasgo que se encuentra bajo una intensa presión selectiva y en la que influyen múltiples variables complejas. Existen dos tipos de melanina, la facultativa que se produce como respuesta a un factor endógeno o exógeno, como es el caso de la luz UVA, provocándose un consecuente oscurecimiento de la piel. Por otra parte la constitutiva, inherente a nuestro cuerpo, cuya expresión se ve modificada por

variaciones en MC1R, así como por ciertas proteínas (Dickkopf 1 del gen DKK1), que provocan diferencias de pigmentación entre distintos sitios del cuerpo (Pathak, 1985; Brenner, 2008). Finalmente, la cantidad y el tipo de melanina varía con la edad y el sexo (Pulker *et al.*, 2007).

El color de la piel está ampliamente relacionado con el origen geográfico (Chaplin, 2004). De hecho, se han encontrado varios SNPs en genes candidatos de pigmentación que se asocian a un fenotipo específico, y a su vez presentan frecuencias distintas a lo largo de diferentes poblaciones como es el caso de los marcadores de ancestralidad (Bouakaze *et al.*, 2009).

Los genes más importantes en cuanto a la determinación del color de la piel son, principalmente MC1R, mencionado con anterioridad por su importancia en el color del cabello; TYR, situado en 11q14-11q21, y que codifica para uno de los enzimas principales en la síntesis de melanina, dado que cataliza la conversión de tirosina a melanina (King *et al.*, 2003); SLC45A2, situado en 5p14.3-5q12.3, implicado en la maduración del melanosoma; SLC24A4, en 15q21.1 cuya proteína es un intercambiador de cationes y actúa de precursor del melanosoma (Pulker *et al.*, 2007). KITLG que codifica para el ligando del receptor de la tirosina kinasa.

Un estudio realizado por Stokowski (2007) muestra que existen variantes en los genes SLC24A5, SLC45A2 y TYR como son rs1426654, rs16891982 y rs1042602 respectivamente, asociadas con el contenido de melanina el cual fue medido por reflectometría en población sur asiática. Tanto SLC24A5 como SLC45A2 muestran un patrón de distribución de frecuencias alélicas entre las poblaciones bastante inusual, con el alelo mutado prácticamente fijado en población Europea, y el alelo ancestral fijado en otros grupos poblacionales (Norton *et al.*, 2007). Lo cual pone de manifiesto importantes signos de selección en población Europea (McEvoy *et al.*, 2006; Voight *et al.*, 2006), indicando que la evolución hacia piel clara tuvo lugar, al menos en parte, independientemente entre Europa y Asia (Hoggart *et al.*, 2003; Han *et al.*, 2008). En un estudio llevado a cabo por Miller, 2007, se investiga la posible asociación entre el SNP, rs642742 cercano a la región 5'UTR del punto de inicio de la transcripción del gen KITLG, y el color de la piel medido por reflectometría, usando para ello un análisis de covarianza. Se demuestra que, individuos con dos alelos ancestrales para este SNP (poblaciones Africanas) presentan un índice de melanina mayor que los individuos que poseen dos alelos mutados (poblaciones Europeas) (Miller *et al.*, 2007).

Dada la elevada complejidad a la hora de cuantificar y calificar objetivamente el color de la piel, y, tratando de evitar los métodos invasivos convencionales (vía HPLC, con una biopsia de la piel), se ha recurrido a métodos espectrofotométricos, que permitan conocer la relación entre la absorción y la reflexión de la piel (Alaluf *et al.*, 2002; Stokowski *et al.*, 2007).

4.5. Aspectos éticos

Éticamente el uso de marcadores genéticos en la predicción de caracteres físicos para ser empleados en casos judiciales está regulado en países como Holanda. En el Reino Unido se han empleado algunos marcadores genéticos como por ejemplo, los asociados a la predicción del cabello pelirrojo (utilizado por el Servicio de Ciencias Forenses de dicho país) (Datos de 2008) (Kayser, 2009). En el caso de España, solo se ha implementado la confidencialidad de datos, el uso obligatorio de consentimientos informados y la consecuente aprobación por un comité ético.

4.6. Consideraciones finales

Trabajando con rasgos físicos que están ampliamente influenciados por el ambiente, es importante tener en cuenta una serie de consideraciones con objeto de minimizar en la medida de lo posible errores durante la asignación de grupos fenotípicos y el tratamiento de los datos. En el caso del cabello, los niveles de melanina pueden variar con el tiempo, dado que disminuye el número y actividad de melanocitos, generando cabellos blancos y grises. La luz UVA, y el agua salada o clorada pueden aclarar el color. Además, hoy en día, la apariencia del cabello se ha visto afectada por modificaciones de tipo cosmético en su morfología y color. Es por esto que resulta especialmente importante la zona en la que se toma la muestra, sugiriéndose la raíz por estar menos expuesta a estos factores ambientales.

Igualmente la piel se ve influenciada por la acción de la luz del sol, así como por el uso de cremas autobronceadoras, o la asiduidad a cabinas de rayos UVA. Esto hace necesario la toma de medidas espectrofotométricas fuera de los meses de verano, y en zonas del cuerpo menos expuestas. Es destacable mencionar que al medir el color con espectrofotometría, se cuantifica no solo la melanina, sino también los carotenos (amarillos), y la hemoglobina que va por las venas (roja), los cuales influyen también en la apariencia externa del color de la piel.

En el caso del ojo, la influencia ambiental afecta este rasgo en menor medida, sin embargo debe considerarse la posibilidad de encontrar cambios en la coloración como respuesta a una enfermedad (síndrome de Horner o la iridocilitis heterocrómica), de forma espontánea en un 10-20% de la población hasta pasada la adolescencia (Carino OB, 1994), o incluso bajo la administración de fármacos en el tratamiento de glaucoma (Stjernschantz, 2002). Además resulta importante considerar para el diseño de estudios de este rasgo, el análisis de otros marcadores genéticos que no están directamente asociados con la pigmentación, y que no están ampliamente estudiados, pero que afectan la apariencia de coloración del iris.

A modo de reflexión es importante considerar que en criminalística la identificación de los rasgos físicos supone un reto adicional, dado que hoy en día las tecnologías permiten la posibilidad de cambiar totalmente la apariencia física de una persona.

Finalmente es meritorio destacar, la necesidad de profundizar en la investigación de estos rasgos físicos antes de que puedan ser aplicados a casos reales de criminalística. También es importante que se realice la validación de estos estudios en distintos grupos poblacionales. Cuando se hayan mejorado las limitaciones en la predicción, es muy probable que el análisis del color de los ojos, piel y cabello sea una realidad en la rutina forense.

Bibliografía

1. Akane, A., Shiono, H., Matsubara, K., Nakahori, Y., Seki, S., Nagafuchi, S., Yamada, M., Nakagome, Y. (1991). Sex identification of forensic specimens by polymerase chain reaction (PCR): two alternative methods. *Forensic Sci Int* 49(1), 81-88.
2. Alvarez-Iglesias, V., Jaime, J. C., Carracedo, A., Salas, A. (2007). Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. *Forensic Sci Int Genet* 1(1), 44-55.
3. Alves-Silva, J., da Silva Santos, M., Guimaraes, P. E., Ferreira, A. C., Bandelt, H. J., Pena, S. D., Prado, V. F. (2000). The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* 67(2), 444-461.
4. Amigo, J., Salas, A., Phillips, C., Carracedo, A. (2008). SPSSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics* 9, 428.
5. Ancans, J., Tobin, D. J., Hoogduijn, M. J., Smit, N. P., Wakamatsu, K., Thody, A. J. (2001). Melanosomal pH controls rate of melanogenesis, eumelanin/phaeomelanin ratio and melanosome maturation in melanocytes and melanoma cells. *Exp Cell Res* 268(1), 26-35.
6. Auton, A. et al., (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res* 19(5), 795-803.
7. Badano, J. L., Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet* 3(10), 779-789.
8. Baeta, M., Nunez, C., Sosa, C., Bolea, M., Casalod, Y., Gonzalez-Andrade, F., Roewer, L., Martinez-Jarreta, B. (2011). Mitochondrial diversity in Amerindian Kichwa and Mestizo populations from Ecuador. *Int J Legal Med* 126(2):299-302.

9. Bamshad, M., Wooding, S., Salisbury, B. A., Stephens, J. C. (2004). Deconstructing the relationship between genetics and race. *Nat Rev Genet* 5(8), 598-609.
10. Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.
11. Bedoya, G. et al., (2009). Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. *PNAS* 103(19), 7234-7239.
12. Biasutti, R. (1959). *Razze E Popoli Della Tierra*. Unione tipografico-Editrice, Torino.
13. Bito, L. Z., Matheny, A., Cruickshanks, K. J., Nondahl, D. M., Carino, O. B. (1997). Eye color changes past early childhood. The Louisville Twin Study. *Arch Ophthalmol* 115(5), 659-663.
14. Blanco-Verea, A., Jaime, J. C., Brion, M., Carracedo, A. (2010). Y-chromosome lineages in native South American population. *Forensic Sci Int Genet* 4(3), 187-193.
15. Borjas, L., Bernal, L. P., Chiurillo, M. A., Tovar, F., Zabala, W., Lander, N., Ramirez, J. L. (2008). Usefulness of 12 Y-STRs for forensic genetics evaluation in two populations from Venezuela. *Leg Med (Tokyo)* 10(2), 107-112.
16. Borsting, C., Sanchez, J. J., Hansen, H. E., Hansen, A. J., Bruun, H. Q., Morling, N. (2008). Performance of the SNPforID 52 SNP-plex assay in paternity testing. *Forensic Sci Int Genet* 2(4), 292-300.
17. Borsting, C., Rockenbauer, E., Morling, N. (2009). Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. *Forensic Sci Int Genet* 4(1), 34-42.

18. Branicki, W., Brudnik, U., Wojas-Pelc, A. (2009). Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype. *Ann Hum Genet* 73(2), 160-170.
19. Branicki, W., Liu, F., van Duijn, K., Draus-Barini, J., Pospiech, E., Walsh, S., Kupiec, T., Wojas-Pelc, A., Kayser, M. (2011). Model-based prediction of human hair color using DNA variants. *Hum Genet* 129(4), 443-454.
20. Brion, M. et al., (2005). Introduction of an single nucleotide polymorphism-based "Major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages. *Electrophoresis* 26(23), 4411-4420.
21. Budowle, B., van Daal, A. (2008). Forensically relevant SNP classes. *Biotechniques* 44(5), 603-8, 610.
22. Bubul, O., Filoglu, G., Altuncul, H., Freire-Aradas, A., Ruiz, Y., Fondevila, M., Phillips, C., Á, C., Kriegel, A. K., Schneider, P. M. (2011). A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. *FSI: Genetic Supplement Series* e500-e501.
23. Butler, J. M., Shen, Y., McCord, B. (2002). The development of reduced size STR amplicons as tools for analysis of degraded DNA. *Forensic Science International: Genetics* 48 (5), 1054-1064.
24. Butler, J. M., Budowle, B., Gill, P., Kidd, K. K., Phillips, C., Schneider, P. M., Vallone, P. M., Morling, N. (2008). Report on ISFG SNP Panel Discussion. *Forensic Science International: Genetics Supplement Series* 1(1), 471-472.
25. Butler, J. M. (2009). *Fundamental of Forensic DNA Typing*. British Library Cataloguing-in-Publication Data.
26. Butler, J. M. (2011). *Advanced Topics in Forensic DNA Typing: Methodology*. Library of Congress Cataloging-in-Publication Data.

27. Cardon, L. R., Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet* 361(9357), 598-604.
28. Carvajal-Carmona, L. G. et al. (2000). Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Am J Hum Genet* 67(5), 1287-1295.
29. Cavalli-Sforza, L., Menozzoni, P., Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.
30. Cavalli-Sforza, L. L., Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 33 Suppl, 266-275.
31. Cook, A. L. et al., (2009). Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci. *J Invest Dermatol* 129(2), 392-405.
32. Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 11(20), 2463-2468.
33. Daugman, J. (2003). The importance of being random: statistical principles of iris recognition. *Pattern recognition* 36(2), 279-291.
34. Dean, M. et al., (1994). Polymorphic admixture typing in human ethnic populations. *Am J Hum Genet* 55(4), 788-808.
35. Devlin, B., Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics* 55 (4), 997-1004.
36. Dowling, D., Friberg, U., Lindell, J. (2008). Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology and Evolution* 23(10), 546-554.

37. Drobnic, K., Borsting, C., Rockenbauer, E., Tomas, C., Morling, N. (2010). Typing of 49 autosomal SNPs by SNaPshot in the Slovenian population. *Forensic Sci Int Genet* 4(5), e125-7.
38. Duffy, D. L., Box, N. F., Chen, W., Palmer, J. S., Montgomery, G. W., James, M. R., Hayward, N. K., Martin, N. G., Sturm, R. A. (2004). Interactive effects of MC1R and OCA2 on melanoma risk phenotypes. *Hum Mol Genet* 13(4), 447-461.
39. Duffy, D. L., Montgomery, G. W., Chen, W., Zhao, Z. Z., Le, L., James, M. R., Hayward, N. K., Martin, N. G., Sturm, R. A. (2007). A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet* 80(2), 241-252.
40. Eiberg, H., Mohr, J. (1996). Assignment of genes coding for brown eye colour (BEY2) and brown hair colour (HCL3) on chromosome 15q. *Eur J Hum Genet* 4(4), 237-241.
41. Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K. W., Hansen, L. (2008). Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet* 123(2), 177-187.
42. Eriksson, N. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet* 6(6), e1000993.
43. Evett, I., Weir, B. (1998). *Interpreting DNA Evidence. Statistical Genetics for Forensic Scientists* Library of Congress Cataloging-in-Publication Data.
44. Fondevila, M., Phillips, C., Naveran, N., Fernandez, L., Cerezo, M., Salas, A., Carracedo, A., Lareu, M. V. (2008). Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur. *Forensic Sci Int Genet* 2(3), 212-218.

45. Franssen, L., Coppens, J. E., van den Berg, T. J. T. P. (2006). Compensation comparison method for assessment of retinal straylight. *Investigative ophthalmology & visual science* 47(2), 768.
46. Freire-Aradas, A., Fondevila, M., Kriegel, A. K., Phillips, C., Gill, P., Prieto, L., Schneider, P. M., Carracedo, A., Lareu, M. V. (2012). A new SNP assay for identification of highly degraded human DNA. *Forensic Sci Int* 6, 341-349.
47. Frost, P. (1994). Geographic distribution of human skin colour: A selective compromise between natural selection and sexual selection? *Human evolution* 9 (2), 141-153.
48. Frost, P. (2006). European hair and eye color: A case of frequency-dependent sexual selection? *Evolution and Human Behavior* 27(2), 85-103.
49. Frost, P. (2007). Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *Am J Phys Anthropol* 133(1), 779-80; author reply 780-1.
50. Frudakis, T. (2003a). Sequences associated with human iris pigmentation. *Genetics* 165(4), 2071-2083.
51. Frudakis, T., Venkateswarlu, K., Thomas, M. J., Gaskin, Z., Ginjupalli, S., Gunturi, S., Ponnuswamy, V., Natarajan, S., Nachimuthu, P. K. (2003b). A classifier for the SNP-based inference of ancestry. *J Forensic Sci* 48(4), 771-782.
52. Frudakis, T., Terravainen, T., Thomas, M. (2007). Multilocus OCA2 genotypes specify human iris colors. *Hum Genet* 122(3-4), 311-326.
53. Frudakis, T. (2008). *Molecular Photofitting. Predicting ancestry and phenotype using DNA*. Inc., E., British Library Cataloguing-in-Publication Data.
54. Fujimoto, A. et al., (2008). A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet* 17 (6), 835-843.

55. Gill, P. (2001). An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int J Legal Med* 114(4-5), 204-210.
56. Glazier, A. M., Nadeau, J. H., Aitman, T. J. (2002). Finding genes that underlie complex traits. *Science* 298(5602), 2345-2349.
57. Gold, D. H., Lewis, R. A. (2004). *Oftalmología*. American Medical Association.
58. Gonzalez-Andrade, F., Roewer, L., Willuweit, S., Sanchez, D., Martinez-Jarreta, B. (2009). Y-STR variation among ethnic groups from Ecuador: Mestizos, Kichwas, Afro-Ecuadorians and Waoranis. *Forensic Sci Int Genet* 3(3), e83-91.
59. Graf, J., Hodgson, R., van Daal, A. (2005). Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum Mutat* 25(3), 278-284.
60. Graf, J., Voisey, J., Hughes, I., van Daal, A. (2007). Promoter polymorphisms in the MATP (SLC45A2) gene are associated with normal human skin color variation. *Hum Mutat* 28(7), 710-717.
61. Graydon, M., Cholette, F., Ng, L. K. (2009). Inferring ethnicity using 15 autosomal STR loci--comparisons among populations of similar and distinctly different physical traits. *Forensic Sci Int Genet* 3(4), 251-254.
62. Green, R. E. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134(3), 416-426.
63. Grimes, E. A., Noake, P. J., Dixon, L., Urquhart, A. (2001). Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Sci Int* 122(2-3), 124-129.
64. Gudbjartsson, D. F. et al., (2008). Many sequence variants affecting diversity of adult human height. *Nat Genet* 40(5), 609-615.

65. Halder, I., Shriver, M., Thomas, M., Fernandez, J. R., Frudakis, T. (2008). A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Hum Mutat* 29(5), 648-658.
66. Han, J. et al., (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet* 4(5), e1000074.
67. Heaton-Armstrong, A. (1995). Eye-witness testimony and judicial studies. *Med Sci Law* 35(2), 93-94.
68. Hoffmann, T. J. et al., (2011). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* 98(6), 422-430.
69. Homer, N. et al., (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4(8), e1000167.
70. Im, S. W., Kim, H. J., Lee, M. K., Yi, J. H., Jargal, G., Sung, J., Cho, S. I., Kim, J. I. (2010). Genome-wide linkage analysis for ocular and nasal anthropometric traits in a Mongolian population. *Exp Mol Med* 42(12), 799-804.
71. Imesch, P. D., Wallow, I. H., Albert, D. M. (1997). The color of the human eye: a review of morphologic correlates and of some conditions that affect iridial pigmentation. *Surv Ophthalmol* 41 Suppl 2, S117-23.
72. Ingman, M., Kaessmann, H., Paabo, S., Gyllenstein, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* 408(6813), 708-713.
73. Ito, S. (2003). A chemist's view of melanogenesis. *Pigment cell research* 16(3), 230-236.

74. Izagirre, N., Garcia, I., Junquera, C., de la Rua, C.,Alonso, S. (2006). A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Mol Biol Evol* 23(9), 1697-1706.
75. Jablonski, N. G.,Chaplin, G. (2000). The evolution of human skin coloration. *J Hum Evol* 39(1), 57-106.
76. Jeffreys, A. J., Wilson, V.,Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature* 316(6023), 76-79.
77. Jobling, M. A.,Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* 4(8), 598-612.
78. Jobling, M. A.,Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nat Rev Genet* 5(10), 739-751.
79. Jobling, M. A., Hurles, M.,Tyler-Smith, C. (2004). *Human Evolutionary Genetics. Orgin, Peoples & Disease*. Garland Publishing.
80. Jones, B. H., Kim, J. H., Zemel, M. B., Woychik, R. P., Michaud, E. J., Wilkison, W. O.,Moustaid, N. (1996). Upregulation of adipocyte metabolism by agouti protein: possible paracrine actions in yellow mouse obesity. *Am J Physiol* 270(1 Pt 1), E192-6.
81. Landsteiner, K. (1990). *The Specificity of Serological Reactions*. Courier Dover Publications.
82. Kayser, M. et al., (2008). Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* 82(2), 411-423.

83. Kayser, M., Schneider, P. M. (2009). DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. *Forensic Sci Int Genet* 3(3), 154-161.
84. Kayser, M., de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet* 12(3), 179-192.
85. Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., Paabo, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell* 90(1), 19-30.
86. Lamason, R. L. et al., (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310(5755), 1782-1786.
87. Lamoreux, M. L., Zhou, B. K., Rosemblat, S., Orlow, S. J. (1995). The pink-eyed-dilution protein and the eumelanin/pheomelanin switch: in support of a unifying hypothesis. *Pigment Cell Res* 8(5), 263-270.
88. Lander, N., Rojas, M. G., Chiurillo, M. A., Ramirez, J. L. (2008). Haplotype diversity in human mitochondrial DNA hypervariable regions I-III in the city of Caracas (Venezuela). *Forensic Sci Int Genet* 2(4), e61-4.
89. Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A., Kayser, M. (2007). Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann Hum Genet* 71 (Pt 3), 354-369.
90. Lettre, G. et al., (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* 40(5), 584-591.
91. Li, J. Z. et al., (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866), 1100-1104.

92. Lin, J. Y., Fisher, D. E. (2007). Melanocyte biology and skin pigmentation. *Nature* 445(7130), 843-850.
93. Lu, T., Pan, Y., Kao, S. Y., Li, C., Kohane, I., Chan, J., Yankner, B. A. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature* 429(6994), 883-891.
94. Liu, F., van Duijn, K., Vingerling, J. R., Hofman, A., Uitterlinden, A. G., Janssens, A. C., Kayser, M. (2009). Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol* 19(5), R192-3.
95. Liu, F. et al., (2010). Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet* 6, e1000934.
96. Lareu, M. V., García-Magariños, M., Phillips, C., Quintela, I., A. Carracedo, Salas, A. (2012). Analysis of a claimed distant relationship in a deficient pedigree using high density SNP data. *Forensic Science International: Genetics* 6, 350-353.
97. Malamud, C. (2005). *Historia De América*. Alianza Editorial.
98. Manning, J. T., Bundred, P. E., Mather, F. M. (2004). Second to fourth digit ratio, sexual selection, and skin colour. *Evolution and Human Behavior* 25(1), 38-50.
99. Mannucci, A., Sullivan, K. M., Ivanov, P. L., Gill, P. (1994). Forensic application of a rapid and quantitative DNA sex test by amplification of the X-Y homologous gene amelogenin. *Int J Legal Med* 106(4), 190-193.
100. Medland, S. E. (2009). Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am J Hum Genet* 85(5), 750-755.
101. Meissner, C., Ritz-Timme, S. (2010). Molecular pathology and age estimation. *Forensic Sci Int* 203(1-3), 34-43.

102. Mengel-From, J., Borsting, C., Sanchez, J. J., Eiberg, H., Morling, N. (2010). Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci Int Genet* 4(5), 323-328.
103. Morón, G. (1971). *Historia De Venezuela*. Italgráfica.
104. Mou, C., Thomason, H. A., Willan, P. M., Clowes, C., Harris, W. E., Drew, C. F., Dixon, J., Dixon, M. J., Headon, D. J. (2008). Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. *Hum Mutat* 29(12), 1405-1411.
105. Need, A. C., Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* 25(11), 489-494.
106. Newton, J. M., Cohen-Barak, O., Hagiwara, N., Gardner, J. M., Davisson, M. T., King, R. A., Brilliant, M. H. (2001). Mutations in the human orthologue of the mouse underwhite gene (uw) underlie a new form of oculocutaneous albinism, OCA4. *Am J Hum Genet* 69(5), 981-988.
107. Niggemann, B., Weinbauer, G., Vogel, F., Korte, R. (2003). A standardized approach for iris color determination. *Int J Toxicol* 22(1), 49-51.
108. Norton, H. L. (2007). Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol Biol Evol* 24(3), 710-722.
109. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature* 456(7218), 98-101.
110. Parra, E. J., Kittles, R. A., Shriver, M. D. (2004). Implications of correlations between skin color and genetic ancestry for biomedical research. *Nat Genet* 36(11 Suppl), S54-60.

111. Prestes, P. R., Mitchell, R. J., Daniel, R., Ballantyne, K. N., Oorschot, R. A. H. v. (2011). Evaluation of the Irisplex system in admixed individuals. FSI: Genetic supplement series 3, e283-e284.
112. Sanchez-Diz, P., Ramos-Luis, E. (2010). Análisis de genomas completos en genética de poblaciones humanas. In Fósiles y moléculas, aproximaciones a la historia evolutiva de Homo sapiens, González-Martín, A., ed. (Salamanca: Memorias de la Real Sociedad Española de Historia Natural. Segunda época, Tomo VIII), 169-201.
113. Pfaff, C. L. (2001). Population Structure in Admixed Populations: Effect of Admixture Dynamics on the Pattern of Linkage Disequilibrium. American Journal of Human Genetics 68, 198-207.
114. Phillips, C. (2005). Using online databases for developing SNP markers of forensic interest. Methods Mol Biol 297, 83-106.
115. Phillips, C., et al., (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1(3-4), 273-280.
116. Phillips, C., Fondevila, M., Garcia-Magarinos, M., Rodriguez, A., Salas, A., Carracedo, A., Lareu, M. V. (2008). Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers. Forensic Sci Int Genet 2(3), 198-204.
117. Phillips, C. (2009). SNP databases. Methods Mol Biol 578, 43-71.
118. Phillips, C. et al., (2009). Ancestry analysis in the 11-M Madrid bomb attack investigation. PloS one 4(8), e6583.
119. Phillips, C. et al., (2011). Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. Forensic Sci Int Genet 5(3), 155-169.

120. Pospiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A., Branicki, W. (2011). Gene-gene interactions contribute to eye colour variation in humans. *J Hum Genet* 56(6), 447-455.
121. Pritchard, J. K., Stephens, M., Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155(2), 945-959. 119.
Pritchard, J., Donnelly, P. (2001). Case-Control Studies of Association in Structured or Admixed Populations. *Theoretical Population Biology* 60, 227-237.
122. Pulker, H., Lareu, M. V., Phillips, C., Carracedo, A. (2007). Finding genes that underlie physical traits of forensic interest using genetic tools. *Forensic Sci Int Genet* 1(2), 100-104.
123. Puri, N., Gardner, J. M., Brilliant, M. H. (2000). Aberrant pH of melanosomes in pink-eyed dilution (p) mutant melanocytes. *J Invest Dermatol* 115(4), 607-613.
124. Purps, J., Geppert, M., Nagy, M., Roewer, L. (2011). Evaluation of the IrisPlex eye colour prediction tool in a German population. *FSI: Genetic Supplement Series* 3 e202-e203.
125. Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., Richter, D. J., Lander, E. S., Altshuler, D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32(1), 135-142.
126. Rigomar R, Arnd M, Green, M. M. (1992). Glossary of Genetics: Classical and Molecular Segen. *Dictionary of Modern Medicine*.
127. Tamarin, R. (2001). *Principles of Genetics*. McGraw-Hill.
128. Rosenberg, N. A., Li, L. M., Ward, R., Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 73(6), 1402-1422.

129. Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., Feldman, M. W. (2002). Genetic structure of human populations. *Science* 298(5602), 2381-2385.
130. Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* 1(6), e70.
131. Salas, A., Phillips, C., Carracedo, A. (2006). Ancestry vs physical traits: the search for ancestry informative markers (AIMs). *Int J Legal Med* 120(3), 188-9; author reply 190.
132. Sanchez, J. J. (2006). A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27(9), 1713-1724.
133. Sanger, F., Nicklen, S., Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74(12), 5463-5467.
134. Sanna, S. (2008). Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* 40(2), 198-203.
135. Sans, M., Weimer, T. A., Franco, M. H., Salzano, F. M., Bentancor, N., Alvarez, I., Bianchi, N. O., Chakraborty, R. (2002). Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. *Am J Phys Anthropol* 118(1), 33-44.
136. Santos, N. P. (2010). Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INDEL) ancestry-informative marker (AIM) panel. *Hum Mutat* 31(2), 184-190.
137. Schiaffino, M. V. (2010). Signaling pathways in melanosome biogenesis and pathology. *Int J Biochem Cell Biol* 42(7), 1094-1104.

138. Schiaffino, M. V., Tacchetti, C. (2005). The ocular albinism type 1 (OA1) protein and the evidence for an intracellular signal transduction system involved in melanosome biogenesis. *Pigment Cell Res* 18(4), 227-233.
139. Seddon, J. M., Sahagian, C. R., Glynn, R. J., Sperduto, R. D., Gragoudas, E. S. (1990). Evaluation of an iris color classification system. The Eye Disorders Case-Control Study Group. *Invest Ophthalmol Vis Sci* 31(8), 1592-1598.
140. Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., Ferrell, R. E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Genet* 60(4), 957-964.
141. Shriver, M. D. (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet* 112(4), 387-399.
142. Shriver MD, K. G. C., Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics*. 1(4), 274-286.
143. Shriver, M. D., Kittles, R. A. (2004). Genetic ancestry and the search for personalized genetic histories. *Nat Rev Genet* 5(8), 611-618.
144. Sobrino, B., Carracedo, A. (2005). SNP typing in forensic genetics: a review. *Methods Mol Biol* 297, 107-126.
145. Spalvieri, M. P., Rotenberg, R. G. (2004). Genomic medicine. Polymorphisms and microarray applications. *Medicina (Buenos Aires)* 64(6), 533-542.
146. Stokowski, R. P. et al., (2007). A genome wide association study of skin pigmentation in a South Asian population. *Am J Hum Genet* 81(6), 1119-1132.
147. Strachen, T., Read, A. P. (1996). Human Molecular Genetics 2. In 2, Kingston, F., ed. Oxford: BIOS Scientific Publishers Ltd, pp. 129-133.

148. Sturm, R. A., Box, N. F., Ramsay, M. (1998). Human pigmentation genetics: the difference is only skin deep. *Bioessays* 20(9), 712-721.
149. Sturm, R. A., Frudakis, T. N. (2004). Eye colour: portals into pigmentation genes and ancestry. *Trends Genet* 20(8), 327-332.
150. Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P., Stark, M. S., Hayward, N. K., Martin, N. G., Montgomery, G. W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the *HERC2* gene determines human blue-brown eye color. *Am J Hum Genet* 82(2), 424-431.
151. Sturm, R. A. (2009). Molecular genetics of human pigmentation diversity. *Hum Mol Genet* 18(R1), R9-17.
152. Sturm, R. A., Larsson, M. (2009). Genetics of human iris colour and patterns. *Pigment Cell Melanoma Res* 22(5), 544-562.
153. Sulem, P. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 39(12), 1443-1452.
154. Sulem, P. et al., (2008). Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet* 40(7), 835-837.
155. Sullivan, K. M., Mannucci, A., Kimpton, C. P., Gill, P. (1993). A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *Biotechniques* 15(4), 636-8, 640-1.
156. Taillon-Miller, P., Piernot, E. E., Kwok, P. Y. (1999). Efficient approach to unique single-nucleotide polymorphism discovery. *Genome Res* 9(5), 499-505.
157. Takamoto, T., Schwartz, B., Cantor, L. B., Hoop, J. S., Steffens, T. (2001). Measurement of iris color using computerized image analysis. *Curr Eye Res* 22(6), 412-419.

158. Teschendorff, A. E. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* 20(4), 440-446.
159. Toyofuku, K., Valencia, J. C., Kushimoto, T., Costin, G. E., Virador, V. M., Vieira, W. D., Ferrans, V. J., Hearing, V. J. (2002). The etiology of oculocutaneous albinism (OCA) type II: the pink protein modulates the processing and transport of tyrosinase. *Pigment Cell Res* 15(3), 217-224.
160. Tully, G., Sullivan, K. M., Nixon, P., Stones, R. E., Gill, P. (1996). Rapid detection of mitochondrial sequence polymorphisms using multiplex solid-phase fluorescent minisequencing. *Genomics* 34(1), 107-113.
161. Program, U. S. D. o. E. G. (2008). SNP Fact Sheet, http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml.
162. Valenzuela, R. K. (2010). Predicting phenotype from genotype: normal pigmentation. *J Forensic Sci* 55(2), 315-322.
163. Vallone, P. M., Butler, J. M. (2004). AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* 37(2), 226-231.
164. Venter, J. C. et al., (2001). The sequence of the human genome. *Science* 291(5507), 1304-1351.
165. Visscher, P. M., Hill, W. G., Wray, N. R. (2008). Heritability in the genomics era concepts and misconceptions. *Nat Rev Genet* 9(4), 255-266.
166. Visser, M., Kayser, M., Palstra, R. J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res* 22(3):446-55.
167. Walsh, S., Lindenberg, A., Zuniga, S. B., Sijen, T., de Knijff, P., Kayser, M., Ballantyne, K. N. (2011a). Developmental validation of the IrisPlex system:

- determination of blue and brown iris colour for forensic intelligence. *Forensic Sci Int Genet* 5(5), 464-471.
168. Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O., Kayser, M. (2011b). IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet* 5(3), 170-180.
169. Walsh, S. (2011c). DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Sci Int Genet* 6(3), 330-40.
170. Weedon, M. N. (2007). A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat Genet* 39(10), 1245-1250.
171. Weedon, M. N. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 40(5), 575-583.
172. Wilson, R. K., Chen, C., Avdalovic, N., Burns, J., Hood, L. (1990). Development of an automated procedure for fluorescent DNA sequencing. *Genomics* 6(4), 626-634.
173. Wright, S. (1949). The genetical structure of populations. *Annals of Human Genetics* 15(1), 323-354.
174. Zhao, L. D. (2011). Effects of DAPT and Atoh1 overexpression on hair cell production and hair bundle orientation in cultured Organ of Corti from neonatal rats. *PLoS One* 6(10), e23729.
175. Zubakov, D. (2010). Estimating human age from T-cell DNA rearrangements. *Curr Biol* 20(22), R970-1.

*Ilustraciones en portada y capítulos: Zamotamay.
Título: "Mapa fenotípico contra el olvido"
Autor: J. C Zamora Tamayo.*